Faculty of Science, Technology, Engineering and Mathematics
M140 Introducing statistics

Unit 3 Earnings

Contents

Introduction

It is a commonplace observation that men earn more than women. (This is often described as a 'gender differential' in pay.) But how much more? And why? These are questions we shall look at in the first section of this unit. Statistics, by its nature, is far better at answering *how much?* questions than at answering *why?* questions, so we shall concentrate on the former.

In Section 1, we consider the problem of how to compare the earnings of men and women. There are two parts to this problem.

- Finding data which adequately summarise the earnings of all the men and women in the country.
- Finding a measure of the difference between men's and women's earnings based on the available data.

Whilst continuing to look at data on earnings, Sections 2 and 3 revise the techniques for presenting and summarising data which you have used in the first two units of the module, and introduce some new ones. In particular, Section 3 introduces another widely used measure of spread: the standard deviation.

In Section 4, you will use Minitab to do further numerical calculations on data and to draw boxplots.

The final section ties together the central themes of Units 2 and 3 by comparing two important measures – the Consumer Prices Index (CPI) and the Average Weekly Earnings (AWE) index. Taken together, these two indices may help us answer the central question of this unit and the previous one: *Are people getting better or worse off?*

In planning your study, you should note that Section 1 is considerably longer than the other sections in the unit.

Also note that you will be guided to the Computer Book at the end of Subsection 3.1 and in Section 4. Like Unit 2, it is better to do the work at those points in the text, although you can leave it until later if you prefer.

1 Gender and earnings

In this section, the plan is to investigate the relationship between gender and earnings. How should we start this investigation?

In Unit 1 you were introduced to an approach to statistical investigations which is summarised in the four stages of Figure 1.

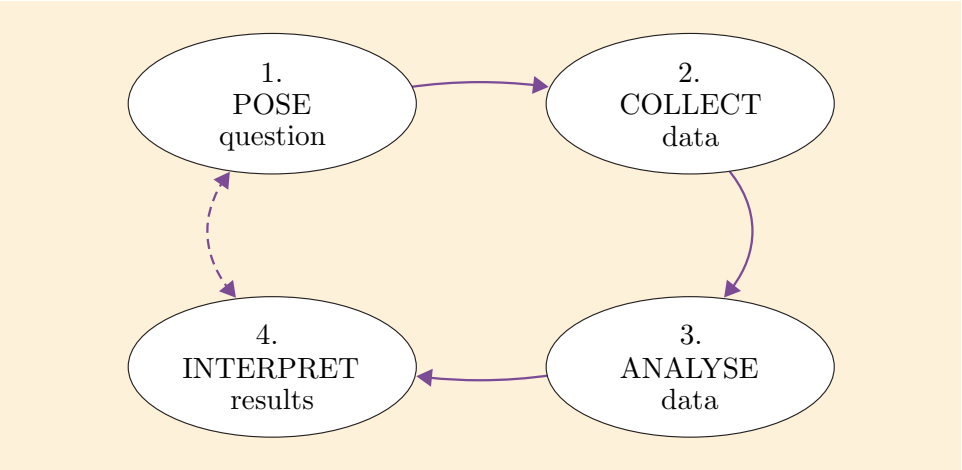


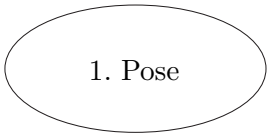
Figure 1 The modelling diagram

So let us follow this plan and start by posing a precise question.

(Throughout Section 1, numbered boxes corresponding to this diagram will be used to point out which of the four stages we are at.)

1.1 Posing the question

Earnings are affected by a great variety of factors apart from gender, including the hours worked, as well as a person's background, training, aptitude and ability. It is useful to illustrate the relationship between earnings and these factors using a diagram like Figure 2.



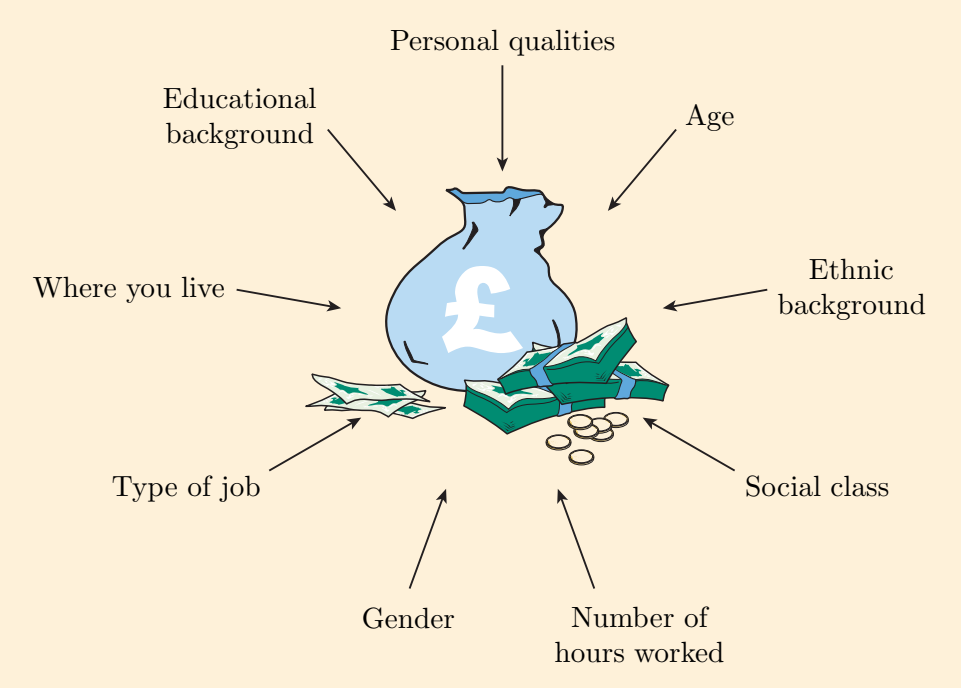


Figure 2 Some of the factors influencing earnings

Note that gender is one of the factors included in Figure 2, but several other factors are included too. Some of these may be more important than gender in determining levels of earnings. One problem in analysing these factors is that they are all interrelated with gender. For example:

- Women may have a different educational background from men, so it is difficult to say whether the differences between men's and women's earnings are

really due to gender or whether they are due to educational background.

- More women than men are part-time workers, so the differences in their earnings may be due to this.
- Some jobs are done predominantly by women, others mainly by men, and earnings vary historically from job to job, so this is another possible reason for differences in earnings.

In social research, problems such as this are common, and it is usually impossible to disentangle completely the effects of the various factors. However, we can go some way towards disentanglement: instead of looking at *all* men and *all* women, we shall concentrate on men and women who are similar with respect to the other variables, so that we compare like with like.

This may not be a straightforward process. A researcher will often know what comparisons he or she would like to make, but available data may not provide the necessary information. There are also the practical constraints of limited time, money, energy and patience to carry out such an investigation. Therefore, *comparing like with like* depends on what is possible and practical.

Determining whether, and to what extent, each of these factors influences earnings, is far beyond the scope of this module. It makes sense to simplify our task by asking the following question.

> *Has there been any change over the last 15 years or so (up to 2012) in the discrepancy between women's and men's earnings?*

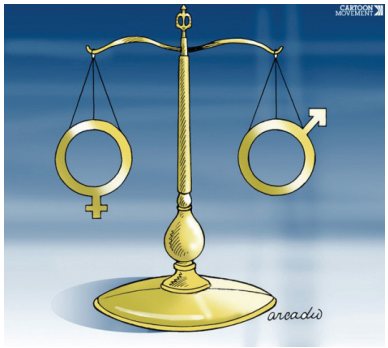
We shall investigate this question in the rest of Section 1.

1.2 Comparing earnings of men and women

Before we can investigate how men's and women's earnings compare, we must do two things:

- Obtain data on the earnings of men and women.
- Decide how to compare their earnings.

The following data come from a UK government website reporting on the Annual Survey of Hours and Earnings (ASHE); in Subsection 1.3 we shall examine this survey and explain exactly why these figures were chosen. (That stage corresponds to the 'collect' box in Figure 1.)



The Equal Pay Act 1970

In 1968, at the Ford car factory in Dagenham, there were 850 women machinists who went on strike for equal pay because they were paid 15% less than men for doing the same work. In the aftermath of their action, the UK parliament passed the Equal Pay Act 1970 that prohibited less favourable treatment between men and women in their terms of employment.

The Act gave an individual a right to the same contractual pay and benefits as a person of the opposite sex in the same employment, where the man and the woman were doing:

- like work; or
- work rated as equivalent under an analytical job evaluation study; or
- work proved to be of equal value.

The Equality Act 2010 simplified anti-discrimination law by consolidating the Equal Pay Act 1970 and numerous other Acts and Regulations into a single piece of legislation.

In 2011, more women than men worked part-time. To attempt to eliminate the effect that this factor has on the relative pay of men and women, the investigation in this section will only concern men and women working full-time; and this only applies to employees being paid adult rates whose pay was not affected by absence in the pay-period for which the data were collected. Although leaving out part-time workers does make it easier to compare like with like, it also means we have omitted many millions of employees, most of them women, from the comparison.

Note that identifying the relevant variables and deciding on appropriate measures of them is an important aspect of any statistical (and often scientific) investigation.

Before looking at the effect of occupation on pay, we shall first consider any difference between the overall earnings of men and women. Table 1 gives the mean gross (that is, before any deductions, such as tax, pension and national insurance, are removed) weekly earnings of adult men and women in full-time employment in the UK in 2011.

(Note that in this unit all the results from the ASHE 2011 are based on the provisional 2011 data, which was available at the time of writing. The revised 2011 data has since become available; however, the changes are minimal and do not change any conclusions made in this unit.)

Table 1 Mean gross weekly earnings of adult men and women in full-time employment (to the nearest pound)

	Women	Men
Mean	515	658

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 1.1)

Having obtained some data, the next step is to decide how to compare the earnings of men and women. From Table 1, it is clear that the mean gross weekly earnings for men is greater than the corresponding figure for women. Two methods of comparing men's and women's mean earnings might spring to mind: subtracting one from the other to find the (numerical) difference, and dividing one by the other to find the ratio.

First, consider the *absolute* numerical difference: this is $\$(658 - 515) = \$143$. So the mean gross weekly earnings of adult men is *$*143 more than the mean gross weekly earnings for women.

Is this a useful way of comparing the earnings of men and women?

Suppose that the mean weekly earnings of men and women had been *$*358 and *$*215, respectively; in this case, the absolute numerical difference would also have been *$*143, as it would have been if the weekly earnings had been *$*1658 and *$*1515. However, an absolute difference of *$*143 would be regarded as of much greater importance in the first case than in the second. So it would be better to know something about the *relative* size of the difference and not just the *absolute* difference.

In Unit 2, ratios rather than absolute differences were used to compare prices. One of the benefits of using ratios is that whatever the unit of measurement – pounds, pence, euros, dollars – the ratio remains the same. Also, because ratios do not depend on the absolute size of the quantities being compared, only on their relative size, ratios calculated at different times can be meaningfully compared. So using relative comparisons makes it possible to extend the investigation to make international comparisons or to make comparisons over time.

So let's use earnings *ratios*. Since the available data in Table 1 are the *mean* gross weekly earnings for men and for women, take the ratio of these means first. The technical term for this is the *earnings ratio at the mean*.

You will use other ratios later.

Earnings ratio at the mean

This earnings ratio at the mean is defined as:

$$\frac{\text{mean earnings of women}}{\text{mean earnings of men}}.$$

An established convention is to take this ratio as the mean women's earnings divided by the mean men's earnings, rather than the other way round.

Example 1 Calculating an earnings ratio at the mean

3. Analyse

For the data in Table 1, the earnings ratio at the mean is

$$\frac{515}{658} = 0.782\,6748 = 0.78 \text{ (to two decimal places).}$$

We shall generally round ratios like this to two decimal places.

In fact, earnings ratios are usually expressed as percentages. Thus, the mean gross weekly earnings of adult women in full-time employment in 2011 was approximately 78% of the mean gross weekly earnings of adult men in full-time employment. (Rounding the ratio to two decimal places corresponds to rounding the percentage to the nearest one per cent.)

Example 1 is the subject of Screencast 1 for Unit 3 (see the M140 website).

In a context where men usually earn more than women, the earnings ratio at the mean will usually be less than one or, as a percentage, less than 100%. The nearer the earnings ratio at the mean is to 100%, the closer are the 'average' earnings of women to those of men.

Table 1 gives the mean gross weekly earnings of adult men and women in full-time employment. This 'compares like with like' to some extent, by avoiding the effects of part-time work and of being paid on non-adult rates. However, there are other factors that affect earnings, like total hours worked, amount of overtime and occupation. Might any of these factors have an effect on the relative earnings of men and women? If they do, then, in order to make a fair comparison, they should be taken into account. How can you find out what effects they have, and take these into account by excluding them from the comparison? We start by looking at hours worked and overtime. You might expect there could be differences between the hours worked and overtime of men and women.

Table 2 Mean weekly hours worked by adult men and women in full-time employment in the UK in 2011 (to one decimal place)

	Women	Men
Normal basic	36.8	38.7
Overtime	0.5	1.5
Total	37.4	40.2

(Source: *Annual Survey of Hours and Earnings*, 2011, Tables 1.9, 1.10 and 1.11)

Note that in Table 2 the normal and overtime average hours do not exactly add up to the total average hours for women, because of a discrepancy caused by rounding.

Activity 1 Hours and overtime

(a) On average, how many hours did women work per week in 2011? (In this context, interpret the phrase 'on average' as 'using the *mean*'.) Was this more or less than the average number of hours worked by men?

(b) On average, did men or women do more overtime per week in 2011, and by how much?

(c) What do you think the effect would be of excluding overtime pay from the mean gross weekly earnings used to calculate the earnings ratio at the mean? Do you think this earnings ratio would increase or decrease if overtime were excluded?

(d) Men and women worked a different number of hours per week on average. Can you suggest a way of eliminating any effect due to this?

Activity 2 Calculating earnings ratios

Table 3 Mean gross weekly and hourly earnings, excluding overtime, of adult men and women

	Women	Men
Mean gross *weekly* earnings excluding overtime (*$*)	509	635
Mean gross *hourly* earnings excluding overtime (pence)	1382	1643

(Source: *Annual Survey of Hours and Earnings*, 2011, Tables 1.2 and 1.6)

In Example 1, the earnings ratio at the mean based on gross weekly earnings, including overtime, was found to be 78%. Use the data in Table 3 to do the following:

(a) Calculate the earnings ratio at the mean based on gross *weekly* earnings excluding overtime.

(b) Calculate the earnings ratio at the mean based on gross *hourly* earnings excluding overtime.

(c) Describe the effect on the earnings ratio at the mean of excluding overtime and using data for hourly earnings instead of weekly earnings.

As you should have seen in part (c) of Activity 2, the longer average working week and the extra overtime worked by men each account for part of the difference between the weekly earnings of men and women. Therefore, to find out whether or not groups of men and women receive equal pay for a similar

amount of work, we need to use gross hourly earnings, excluding overtime (when available), for a fairer comparison.

1.3 The Annual Survey of Hours and Earnings

The data which we use to compare men's and women's earnings in this unit are taken from a government survey called the Annual Survey of Hours and Earnings (ASHE). This annual survey has been carried out by the Office for National Statistics (ONS) each April since 2004. Before that there was a similar survey called the New Earnings Survey, carried out from 1970 to 2003.

The main purpose of the survey is to provide information on patterns of earnings and paid hours for employees within industries, occupations and regions. The survey results are used by the UK government and many other organisations, both public and private. Among other uses, ASHE data are used by the UK government to help set minimum wage levels, and (as we are doing in this unit) to investigate the gender pay-gap.

Currently, ASHE collects information on about 180 000 men and women who are members of the Pay-As-You-Earn (PAYE) scheme. Thus the survey does not cover the self-employed nor, of course, unemployed people. (The survey also omits people whose employer does not employ anyone who earns above the (low) threshold that requires the business to register for PAYE; however, ONS investigations indicate that this omission does not cause a major bias.)

To collect the data, the employer of each selected employee is contacted. The employee is identified by name and National Insurance number and the employer is required, by law, to provide information relating to that employee's earnings during a specified week.

The data collected include (among other things) the following items of information concerning the employee:

- total earnings for the pay-period including the specified week
- location of workplace
- occupation
- information concerning normal basic hours, overtime earnings and hours, bonus payments, pension contributions and length of pay-period.

The employee's age and gender are also recorded from other government sources. This information is analysed by ONS, along with information on national numbers of employees in specific groups taken from another government survey, the Labour Force Survey. (It is this extra information that, among other things, gets round the issue of the omission of employers whose employees all fall below the PAYE threshold.) The results are published online on the ONS website.

1.4 Averages: the mean or the median?

So far you have used the mean when comparing the levels of earnings of men and women. But there are other measures for summarising data: the median, in particular. Would the results of the investigation have been the same using *median* earnings of men and women? You already know from Unit 2 that, as measures of location, the median and the mean differ in important ways – for instance, they differ in their resistance to extreme values in the data. But here let us look at a different aspect, connected to the *skewness* of the data.

The mean earnings of any group of people may be thought of as the 'average' of the earnings of all the people in that group. The median earnings may be thought of as the earnings of the 'middle-income person', or more precisely (as income also includes things other than pay), as the person on middle earnings; that is, if you listed the people in order of their earnings, then the person halfway up the list (and halfway down) gets the median earnings. So mean earnings and median earnings are different ways of measuring the 'middle' level of earnings of the group.

Table 4 gives values of the median and mean gross weekly earnings including and excluding overtime, and the median and mean gross hourly earnings excluding overtime for adult men and women in full-time employment in 2011.

Table 4 Median and mean earnings for men and women in 2011

	Median		Mean	
	Women	Men	Women	Men
Gross weekly earnings incl. overtime ($)	440	538	515	658
Gross weekly earnings excl. overtime ($)	432	509	509	635
Gross hourly earnings excl. overtime (p)	1174	1311	1382	1643

(Source: *Annual Survey of Hours and Earnings*, 2011, Tables 1.1, 1.2 and 1.6)

The *earnings ratio at the median* is defined in a similar way to the earnings ratio at the mean.

Earnings ratio at the median

The earnings ratio at the median is defined as:

$$\frac{\text{median earnings of women}}{\text{median earnings of men}}.$$

Activity 3 The earnings ratio at the median

3. Analyse

(a) Calculate the earnings ratio at the median using the data in Table 4 for each of the following: gross weekly earnings including overtime, gross weekly earnings excluding overtime, gross hourly earnings excluding overtime.

(b) Compare the three earnings ratios at the median that you calculated in part (a) with the corresponding earnings ratios at the mean (which were calculated in Example 1 and Activity 2). This latter set of values is 78%, 80% and 84%, respectively. What do you notice?

Activity 3 showed that in each considered case the earnings ratio at the median is greater than the earnings ratio at the mean. Remember, the nearer any earnings ratio is to 100%, the closer the earnings of women are to those of men. So the relative 'gap' between median earnings is less than the relative gap between mean earnings.

Looking again at Table 4, it can also be seen that, for both men and women, the median earnings figure is less than the mean earnings figure. In fact, for earnings data, it is generally true that the median is smaller than the mean. Why should this be so?

Here is an example of a typical *distribution* of earnings; that is, how earnings vary between employees. Imagine a small manufacturing company. The earnings of the majority of the employees will probably not be very different from one another: maybe some will earn as much as twice the amount that others do, but not much more. However, there will almost certainly be one or two senior managers who earn very much more. This hypothetical distribution of earnings is illustrated in Figure 3.

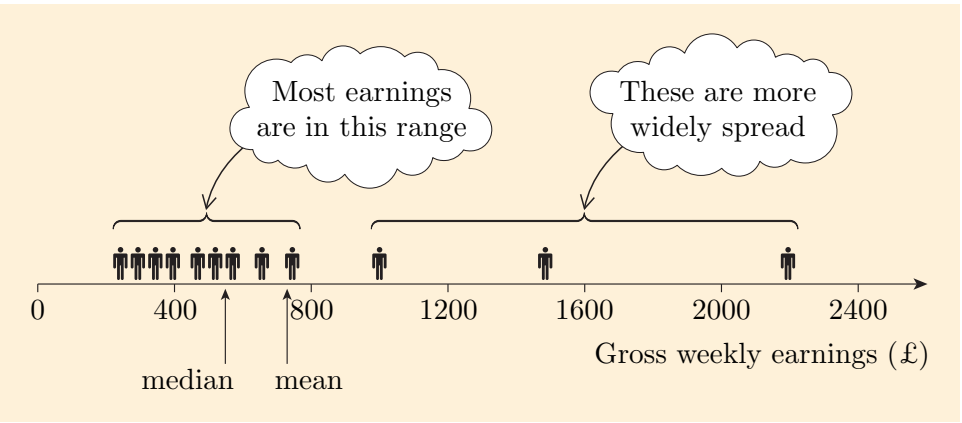


Figure 3 The distribution of earnings in a small (imaginary) company

So the earnings of the majority of employees will be fairly closely grouped, but there will be a few who earn much more. This is the case for earnings in general. This distribution of earnings is the primary reason for the phenomenon that median earnings are generally lower than mean earnings.

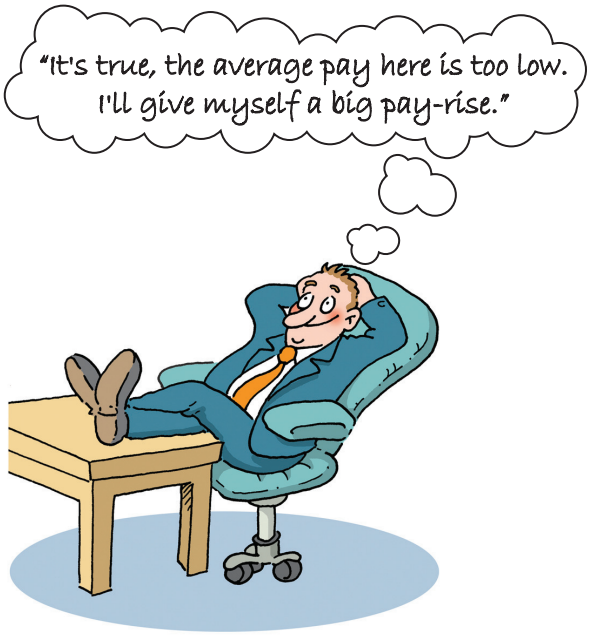
To see why this is, imagine that the gross weekly earnings of the highest-paid employee in the imaginary company goes up from about $2200 (as in Figure 3) to $3000. What effect would that have on the median earnings?

The median would not change (if the other pay levels remain unchanged), because however large the highest-paid employee's pay becomes, the middle values remain the same. The mean, however, is the total earnings of the

12 employees divided by 12. So, as the highest-paid employee's pay increases, the total earnings of the 12 employees increases and hence the mean rises. Increasing the highest-paid employee's pay has the effect of dragging up the mean, even if everyone else's pay remains the same.

Since, in general, higher earnings are much more widely spread than lower earnings, this phenomenon of the mean being greater than the median is to be expected when examining earnings data. Further, the more widely spread the higher earnings are, compared with the lower earnings, the greater the difference will be between the mean and the median.

For the earnings data in Table 4, the mean is greater than the median for both men and women. This reflects the fact that higher earnings are more widely spread than lower earnings for both men and women. Also, in Activity 3, you found that the earnings ratio at the median is greater than the earnings ratio at the mean in each case, so the relative gap between median earnings is less than that for mean earnings. This is due to higher earnings being more widely spread, compared with lower earnings, for men than for women.



Decisions about which measure to use have implications for how the data are interpreted. Since mean earnings are generally higher than median earnings, a trade union official might use median earnings to support an argument that average pay is low. On the other hand, the employer might use the mean to argue the opposite. As the relatively large earnings of a minority can have a marked effect on mean earnings, whereas the median is unaffected by a few extremely large values, so there is a case for regarding the median as more representative of earnings in general than the mean. For the rest of the investigation into the earnings of men and women here, we shall use the median rather than the mean.

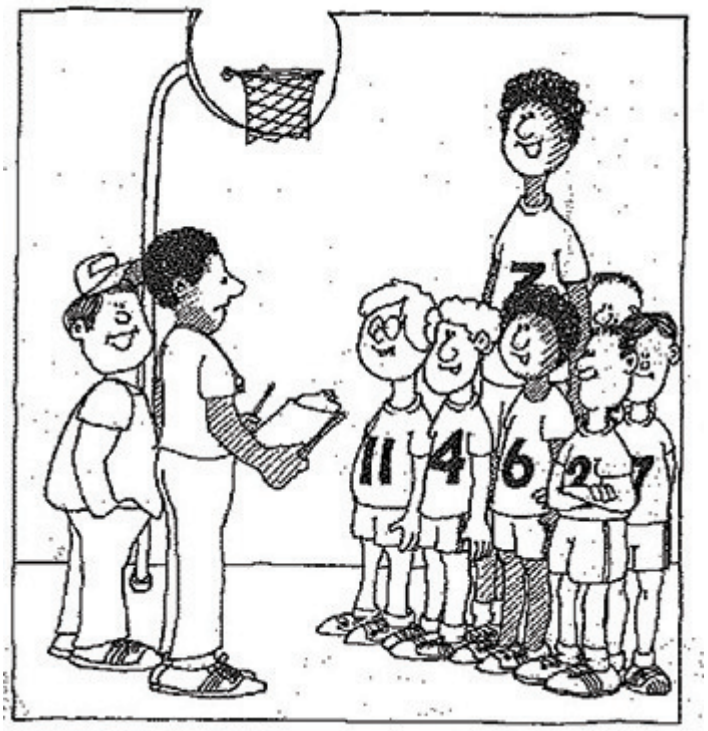
There is, however, nothing *unique* to earnings data that leads to the median being less than the mean. Since this arises from the higher values being more widely spread out than the lower values, it will occur in any data with that characteristic. You know from Units 1 and 2 that data of this kind, where the higher values are more widely spread out than the lower values, are said to be right-skew. So in right-skew data, the median will generally be less than the mean.

What about left-skew data? Well, everything works in the opposite direction. In left-skew data, the lower values are more widely spread out, compared to the higher values, and this leads to the median generally being greater than the mean. These characteristics are summarised below.

Skewness, the median and the mean

In *right-skew* data, the median is generally *less* than the mean.

In *left-skew* data, the median is generally *greater* than the mean.



'Should we scare the opposition by announcing our mean height or lull them by announcing our median height?'

The pay parade

The following cartoon appeared in a journal article in 1994. Imagine a parade of all the workers in the UK, in which everyone's height is proportional to their weekly earnings: so a person earning an average (that is, mean) wage is of mean height. The shortest, that is the lowest-paid, is first in the parade, the tallest, that is the highest-paid, is last. Suppose the parade takes an hour with everyone moving at the same speed. For the first 25 minutes all you will see are very short people – nearly ten million people less than four feet tall. Only in the last 25 minutes do you see people of average (mean) height, followed by a few giants: government ministers who are nine metres tall and heads of companies who are as tall as a skyscraper. (This idea of a 'pay parade' dates back at least to a 1971 publication by the Dutch economist Jan Pen. The figures in the text, but not the cartoon itself, have been updated to relate to the UK position in 2011.)

Activity 4 The pay parade: calculations and interpretations

Suppose that the parade, described above, begins at 10 am and ends at 11 am.

(a) At what time would a person of median earnings pass by?

(b) According to the description of the pay parade, at what time does a person earning the mean wage pass by? How does this relate to the type of skewness in these income data? What percentage of people earn less than the mean wage?

Activity 5 The pay parade: using images

Think about the cartoon image itself: what were some of your reactions to it? Were there any aspects that confused you, or where you felt you were being misled?

Though the cartoon provides a powerful image, some images can be misleading! In the cartoon, a person earning twice the average wage is drawn twice as tall as a person earning the average wage. However, that person is also drawn twice as wide – the tall people are not tall and thin – so the *area* of the cartoon taken up by a person earning twice the average wage is *four* (2×2) times the area taken up by a person earning the average wage. And, in practice, a reader may well interpret a person in the cartoon as a figure occupying a *volume* in space. So the impression received is of a figure *eight* $(2 \times 2 \times 2)$ times as large. Thus, the effect of the cartoon is to exaggerate the differences in earnings of different people.

Unfortunately, many published diagrams make use of area or volume to exaggerate the visual effect of points they are trying to make. Look out for this whenever you see diagrams used to support an argument.

The cartoon is based on the idea that greater height corresponds to greater income. This choice has quite strong psychological overtones to do with cultural norms of 'stature', 'importance', and so on; it is a far from neutral image.

Consider the impact of a redrawn cartoon where the key image was a person with their hand outstretched: the larger the salary, the longer the arm.

One reason you may be able to orientate yourself with regard to the cartoon as it stands, is that you have plenty of experience of the distribution of people's heights and you can use this to interpret the image. Most importantly, there is no scale, other than the notion of 'average height'.

Let's turn away from the pay parade and summarise the investigation so far. Several factors that affect earnings have been taken into account. Since more women than men work part-time, the investigation was restricted to full-time workers. Only workers on adult pay rates were considered. Since men work more overtime on average than women, overtime was excluded. Since the normal basic working week is slightly longer on average for men than for women, the average hourly earnings of men and women were compared instead of the average weekly earnings. Since a few well-paid individuals can strongly influence the mean but not the median, the median was used for comparisons. This is summarised in Table 5.

Table 5 Adjustments made in order to compare 'like with like'

Perceived problem	Proposed solution
More women than men work part-time.	Look only at full-time workers on adult pay rates.
Men work more overtime.	Exclude overtime.
Men work a longer basic working week.	Compare hourly earnings.
A few highly-paid individuals can seriously influence the mean.	Compare median earnings.

Even after taking all these factors into account, the earnings ratio at the median for 2011 was 90%. So it appears that adult women working full-time receive a median hourly rate that is only about 90% of the median hourly rate for men (see Activity 3 earlier in the subsection). Does this mean that women and men are not receiving equal pay for equal work? Or are there other aspects that have not yet been taken into account?

Perhaps the most important factor influencing pay that has not yet been considered, is actual occupation; one aspect of this is briefly investigated later in the unit.

We have seen that looking at data on hourly (rather than weekly) pay has advantages. However, particularly when looking at individual occupations, there are disadvantages too. In many occupations, there is no paid overtime: pay is based on a nominal number of contracted hours and is fixed regardless of the number of hours actually worked, so that the published figure for hourly earnings may bear little relationship with reality. This is the case in many professions.

In secondary-school teaching, for example, the basic working week for full-time workers according to the results of the ASHE is a little over 32 hours. However, this relates only to 'directed' hours of work, and teachers are expected (and indeed required, under their contracts of employment) to carry out other duties, such as much of their lesson preparation and marking, outside these hours. Indeed, many surveys have established that secondary-school teachers work, on average, many more hours in a week than the 32 in their basic working week, generally without any extra pay. But ASHE data on hourly pay do not take into account these extra working hours. In such occupations, ASHE figures for hourly

earnings are fairly meaningless and should certainly not be used for comparisons with earnings in other occupations.

Taking all these factors into account, the module team settled for using data on weekly earnings excluding overtime for most comparisons in M140.

Another important aspect is that, so far, the only features of the pay distribution that we have considered are its mean and median. In the next subsection, other features are considered. In terms of the pay parade, we ask whether there are ways in which the numbers of men and women vary towards the start and end of the parade, rather than in the middle.

1.5 Deciles

You have seen data on median and mean pay from ASHE, for various groups of workers. But the ASHE results do not provide only means and medians. For instance, Table 6 shows the data presented in the 2011 ASHE report on gross weekly pay excluding overtime, for male and female workers separately. (The data are for workers on adult rates of pay, whose pay was not affected by absence.) To begin with, we'll concentrate on the data for male workers.

Table 6 Weekly pay excluding overtime (rounded to the nearest *$*) for full-time employee jobs in the UK, 2011

Percentile	Female	Male
10	253	284
20	296	340
25	316	364
30	335	390
40	380	446
60	497	581
70	574	676
75	619	738
80	671	813
90	820	1083
Median	432	509
Mean	509	635

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 1.2)

Well, you know what the median and mean are, but what about all these 'percentiles'?

You are already familiar with the idea of the quartiles of a batch of data 'cutting' the batch into quarters. That is, when the numbers in the batch are sorted into order, one quarter of the values are below the lower quartile, and one quarter are above the upper quartile. Percentiles are a related idea. According to the data for male workers in Table 6, the percentile labelled 10 for these data is *$*284. This quantity is usually called the *10th percentile*, and what it means (in this case) is that 10% of men in 2011 earned less than the 10th percentile, that is, less than *$*284 a week. Similarly, for these data, the 20th percentile is *$*340, which means that 20% of men earned less than *$*340 per week.

In general the nth **percentile** is the number such that $n\%$ of the values in the batch fall below it.

Activity 6 Percentiles in the ASHE

(a) Most of the percentiles in Table 6 correspond to a percentage that is a multiple of 10. However, two of them, the 25th and the 75th percentile, are not multiples of 10. Why do you think these two percentiles are included? What is the other name for these two percentiles?

(b) The table does not give the zeroth or the 100th percentile – these would be the extremes (the maximum and minimum) of the dataset. Apart from that, the column of the table labelled 'Percentile' gives all the percentiles corresponding to a percentage that is a multiple of 10, *except* the 50th percentile. Why is that one omitted? (Hint: it appears elsewhere in the table.)

So Activity 6 explains that some percentiles are the same as quantities for which there is another name. The 25th percentile is the lower quartile, the 75th percentile is the upper quartiles, and the 50th percentile is the median.

Deciles

'Decile' comes from the Latin word 'decem', which means 'ten'.

Percentiles for which the corresponding percentage is a multiple of 10 also have another name: **deciles**. This is because they divide up the batch of data into tenths.

One tenth of the values are below the 10th percentile, one tenth are between the 10th and the 20th percentiles, and so on. In particular, the 10th percentile is called the **lowest decile** and the 90th percentile is called the **highest decile**.

The lowest and highest deciles of a batch of data help us to investigate what are known as the *tails* of its distribution. Just as the quartiles cut off the extreme quarters, or 25%, at either end, the highest decile cuts off the top 10% whilst the lowest decile cuts off the bottom 10%. Figure 4 illustrates these and other key positions we have seen so far (excluding the mean).

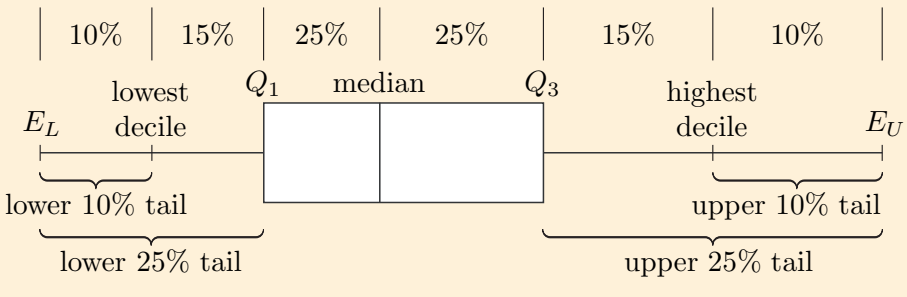


Figure 4 Extremes, lowest and highest deciles, quartiles and median

Example 2 Interpreting some special percentiles

If you look at the figures for male workers in Table 6, you can see that the median earnings of males on adult rates in full-time employment (whose pay was not affected by absence) was *$*509 (rounded to the nearest pound). This means that, in a week, 50% of this group of men earned more than *$*509 whilst 50% earned less than *$*509. The upper and lower quartiles for men, given in Table 6, are 738 and 364, so 25% of this group earned less than *$*364, whilst 50% earned

4. Interpret

between *$*364 and *$*738, and 25% earned more than *$*738. The extra information we obtain by looking at the highest decile, 1083, and the lowest decile, 284, is that only 10% of this group earned more than *$*1083 and only 10% earned less than *$*284. Recall that we are not told what the extreme values are.

You learned in the last unit why a boxplot is a useful diagrammatic representation of a batch of data. It could therefore be informative to draw a boxplot of the data in Table 6. However, we cannot draw a boxplot exactly like those in Unit 2 since there we marked the extremes of the batch, E_L and E_U, on the boxplot and we do not know the extremes of this batch.

It is quite common, when dealing with large batches of data from surveys like the ASHE, not to know the extreme values. So we often draw a boxplot which extends only from the lowest decile to the highest decile. A boxplot like this is called a **decile boxplot**.

Example 3 A decile boxplot

Figure 5 shows a decile boxplot of the weekly earnings of the group of adult male full-time employees discussed in Example 2. Arrowheads are used at the end of the whiskers, to remind us that these points represent the highest and lowest deciles and *not* the extremes.

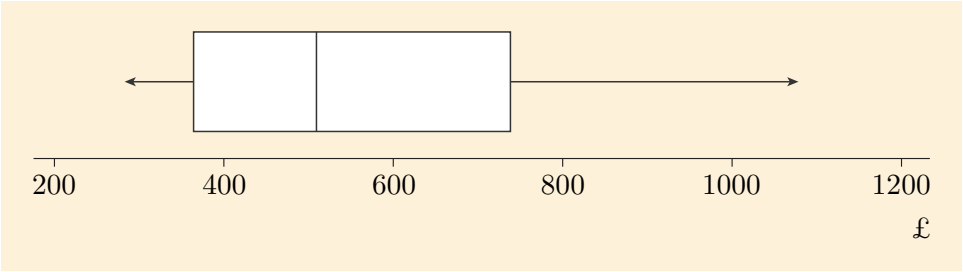


Figure 5 Decile plot of weekly earnings for male full-time employees

---

Example 3 is the subject of Screencast 2 for Unit 3 (see the M140 website).

When you have a batch of data, large or small, you will be, for many purposes, most interested in investigating the main body of the data; that is, the part in the middle of the distribution that contains the most typical data values. For example, you might often look just at the median and the quartiles. However, studying the more extreme values that are far from the median – the **tails** of the distribution – is often also important, as you will see later in this module. This method of describing the tails by giving the highest and lowest deciles can be very useful. Indeed, it is common to go further out into the tails and look for the points that separate off the top and bottom 5% of the values, or further still to *cut off* the top and bottom 2.5%, 1% or 0.5%.

As with the deciles and quartiles, the end parts of the distribution that are cut off are called the 5% tails, or the 2.5% tails, or the 1% tails, or the 0.5% tails. Some of these tails are shown in Figures 6 and 7.

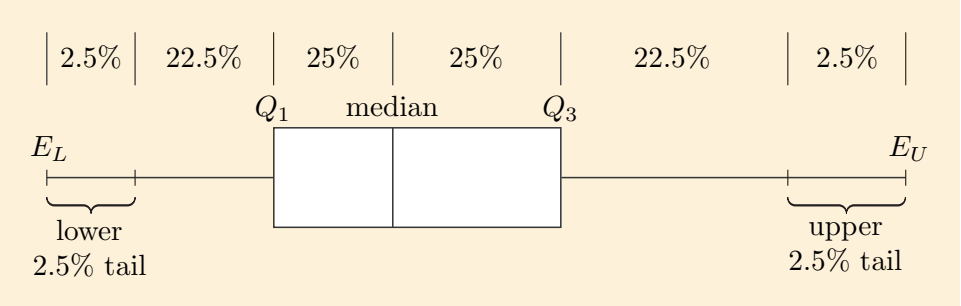


Figure 6 Lower and upper 2.5% tails

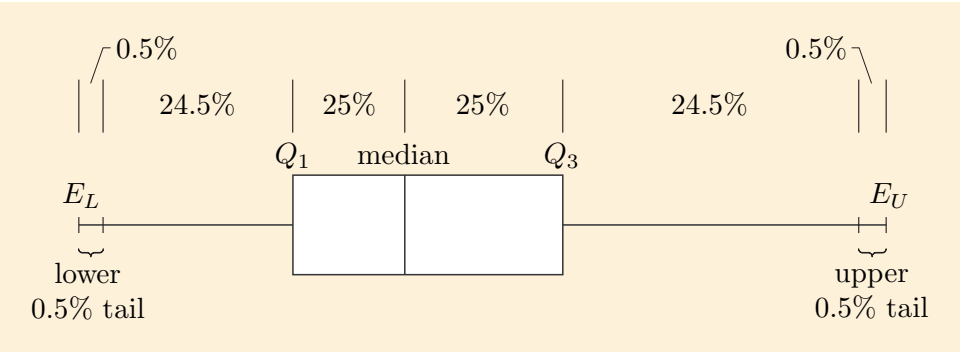


Figure 7 Lower and upper 0.5% tails

In Subsection 1.6 we shall return to our investigation of men's and women's earnings, but first, here is an activity to test your comprehension of the ASHE data that we shall be using, and of the various ways you have met to describe the data.

Activity 7 Extracting information from a table of percentiles

These questions concern the group of male full-time employees on adult rates included in Table 6. ASHE, the source for these data, also informs us that the number of employees in this group (in the UK as a whole, not in the actual sample) was estimated as 10 652 000.

(a) What percentage of the men in this group earned more than *$*1083 in that week?

(b) What percentage of the men in this group earned between *$*284 and *$*509 in that week?

(c) Approximately how many men in this group earned less than *$*364 in that week?

1.6 Earnings ratios across the distribution

1. Pose

Let us now return to our investigation comparing men's and women's earnings. We have dealt with many aspects of the need to compare like with like. As mentioned earlier, there is a need for compromise in this regard, and Table 5 (in Subsection 1.4) re-emphasises this. That table does not include several factors that Figure 2 (in Subsection 1.1) identifies as affecting earnings (educational background is one example). Also, we have not yet considered how to take into account different occupations. However, other variables have been taken into account; for instance:

- age (to the limited extent of considering only those on adult rates of pay)
- hours worked (full-time or part-time)
- absentees (those whose pay for the survey pay-period was affected by absence).

In order to compare like with like we shall, in this subsection, compare only employees who are:

- on adult rates of pay
- full-time
- non-absentees.

We shall also exclude overtime pay. The question of whether to use weekly or hourly pay was also raised in the text below Table 5 – there are pros and cons, but for the rest of this section we shall use weekly pay.

Table 6 thus contains the relevant data for making the comparison for 2011. For convenience, here is the table again, but giving only the data that will be most relevant for our purposes. These are the figures for the highest decile (90th percentile), the upper quartile (75th percentile), the median (50th percentile, though it is not labelled in that way in Table 6),the lower quartile (25th percentile), and the lowest decile (10th percentile).

Table 7 Weekly pay excluding overtime (rounded to the nearest *$*) for full-time employee jobs in the UK, 2011

	Female	Male
Highest decile	820	1083
Upper quartile	619	738
Median	432	509
Lower quartile	316	364
Lowest decile	253	284

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 1.2)

You have already calculated (in Activity 3 of Subsection 1.4) the earnings ratio at the median for these data: it was 85%. However, that compares only the 'average workers' in the middle of the earnings distribution. Just as for the median (and the mean), earnings ratios at the quartiles and the deciles can be defined.

Earnings ratios at the quartiles and deciles

The **earnings ratio at the lower quartile** is:

$$\frac{\text{lower quartile earnings of women}}{\text{lower quartile earnings of men}}.$$

The **earnings ratio at the lowest decile** is:

$$\frac{\text{lowest decile earnings of women}}{\text{lowest decile earnings of men}}.$$

The **earnings ratio at the upper quartile** is:

$$\frac{\text{upper quartile earnings of women}}{\text{upper quartile earnings of men}}.$$

The **earnings ratio at the highest decile** is:

$$\frac{\text{highest decile earnings of women}}{\text{highest decile earnings of men}}.$$

3. Analyse

4. Interpret

Activity 8 Calculating earnings ratios

Calculate the earnings ratios at the upper quartile, the lower quartile, the highest decile and the lowest decile, for the workers represented in Table 7. (Round your answers to the nearest one per cent.)

To make sense of all these ratios, it is helpful to put them in the order that the various percentiles would come in a batch of data, as in Table 8.

Table 8 Earnings ratios at a range of percentiles

Percentile	Earnings ratio
Highest decile	76%
Upper quartile	84%
Median	85%
Lower quartile	87%
Lowest decile	89%

At the bottom end of the distribution the earnings ratio is rather higher than at the top end, and the earnings ratio at the highest decile is quite a lot lower than the others. However, the figures indicate that the earnings ratio does not vary too much across the range from the lowest decile to the highest decile. As an overall summary of these figures we can reasonably say that the earnings ratio is about 85%.

In Figure 8, the decile boxplots for the earnings of both the men and women in Table 7 have been drawn on the same scale.

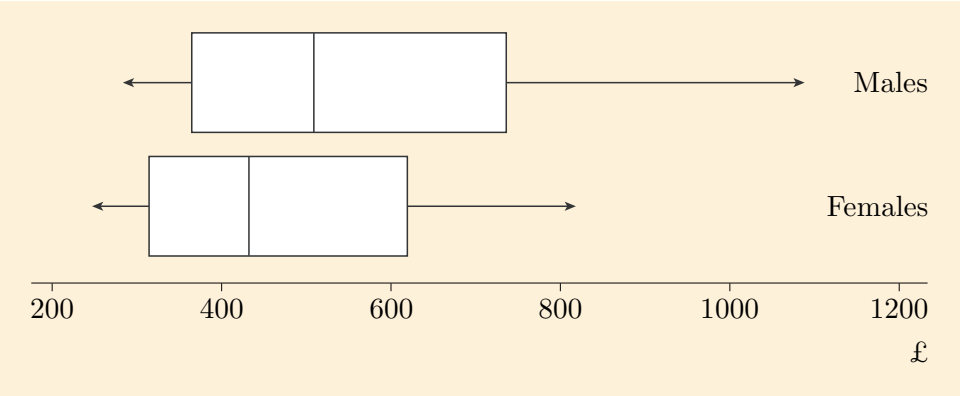


Figure 8 Decile boxplots comparing earnings of men and women

Although there is some degree of overlap between these two boxplots, it is clear that the location of the distribution for males exceeds that for females by quite a margin. Each of the five values plotted (median, quartiles, deciles) for the males is quite a distance to the right (higher on the scale) than the corresponding figure for the females. That is, there still remained a considerable difference between the pay levels for men and women in the UK in 2011.

1.7 Has the 'gap' between men's and women's earnings been closing?

Now that we have a method of comparing earnings, we can return to our question about how the gap between men's and women's pay has changed over the last 15 or so years. We shall return from looking at weekly earnings, to looking at hourly earnings – this is simply because hourly earnings (excluding overtime) for previous years are easier to find! Table 9 shows the median gross hourly earnings, excluding overtime, for adult employees for the years 1997 to 2011.

1. Pose

Table 9 Median gross hourly earnings excluding overtime (in pence)

Year	Women	Men
1997	694	840
1998	722	874
1999	758	907
2000	783	935
2001	823	984
2002	867	1026
2003	904	1058
2004	937	1096
2005	982	1129
2006	1023	1171
2007	1048	1197
2008	1092	1250
2009	1139	1297
2010	1169	1300
2011	1174	1311

(Source: *Patterns of Pay*, 1997 to 2011 ASHE Results, 2011, Table 1)

This table takes the data right back before the ASHE was first carried out in 2004. For data before that, the source is the New Earnings Survey, the predecessor of ASHE, but ONS statisticians have adjusted the numbers to make the basis of comparison between the surveys more accurate. Also, the way the ASHE data were calculated changed in 2006, although the effects of these changes on the earnings ratio is not thought to be too large. These effects will be ignored in what follows, although it means we are not entirely comparing like with like.

Activity 9 Changes in the earnings ratio over time

(a) Calculate the earnings ratio at the median for each year in Table 9.

(b) How has the earnings ratio changed since 1997?

(c) On the evidence of your calculations, would you say that gender inequalities in earnings have widened, narrowed or stayed the same between 1997 and 2011?

3. Analyse

In conclusion, certainly there have been changes in the earnings ratio at the median over the 15-year period considered. Over this period, the median pay of female employees did move closer to that of male employees, on an hourly basis at least. But the increase, while steady, has been rather slow, and on this measure, women in 2011 were still earning rather less than men. But we have not been able to investigate all the possible reasons for this discrepancy, and indeed several of the reasons cannot be investigated using ASHE data alone.

4. Interpret

The next subsection contains a brief investigation of one other aspect of the difference between men's and women's earnings: the effect of different occupations.

1.8 Further investigations into gender and earnings

1. Pose

Figure 2 (in Subsection 1.1) pictured many factors, in addition to gender, that influence earnings. It is quite possible that it is really these other factors, such as educational background and type of job, which are the basic causes of the difference between men's and women's earnings. So in this subsection we shall briefly consider how an investigation into the influences of such other factors might proceed.

2. Collect

For example, one possibility is that the lower earnings of women are due largely to the types of job that they do. The ASHE provides information on the earnings of people in different occupations, and these are the kind of data that are needed to investigate this possibility. As an illustration of one method of starting such an investigation, the distribution of the earnings of men and women in public and private sector jobs is summarised in Table 10. The table excludes some occupations that are not classified as being in either the public sector or in the private sector.

Table 10 Distribution of gross weekly earnings (excluding overtime) of full-time adults in the public and private sectors, 2011 (rounded to the nearest *$*)

	Public		Private	
	Men	Women	Men	Women
Highest decile	1065	844	1095	791
Upper quartile	781	675	709	543
Median	588	510	480	372
Lower quartile	429	377	347	282
Lowest decile	341	307	274	235

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 13.2)

Activity 10 Comparing earnings ratios

Use the data in Table 10 to calculate the earnings ratio at each point for both public sector and private sector occupations. (Round your answers to the nearest one per cent.)

3. Analyse

Compare these ratios with the figures for all full-time workers, calculated in Activities 3 and 8, which were: highest decile 76%, upper quartile 84%, median 85%, lower quartile 87% and lowest decile 89%. What do you notice?

We can also draw decile boxplots for all four of the batches of data described by Table 10.

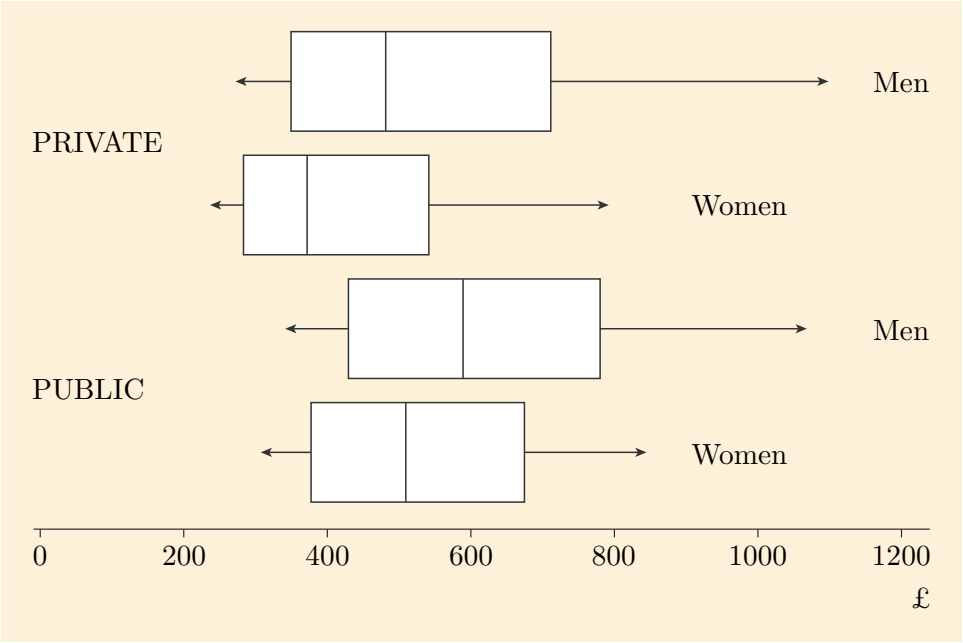


Figure 9 Decile boxplots comparing male and female earnings in the private and public sectors

Comparing the boxplots in Figure 9, you can see that, for both men and women (except the highest-paid men), earnings are lower in the private sector than in the public sector, but the difference is markedly greater for women than for men.

Calculating the earnings ratios of public and private sector employees has not, by itself, given us any clear evidence about the relationship between types of job and the comparison of women's and men's earnings. Now, the public and private categories are themselves composed of a variety of occupations, but, as can be seen from the examples in the next activity, even if we look at more specific occupations, the only consistent pattern that emerges is that women earn less than men.

Activity 11 Comparing earnings ratios for different occupations

Table 11 gives the median and quartiles (rounded to the nearest pound) of the distribution of the gross weekly earnings of full-time adult men and women in a variety of occupations.

Table 11 Gross weekly earnings for a variety of occupations

Occupation:	Sales and retail assistants			Secondary education professionals			Kitchen and catering assistants			Managers and directors in retail/wholesale		
	M	W	R	M	W	R	M	W	R	M	W	R
Upper quartile	352	312		852	801		297	283		698	497	
Median	289	260		737	699		247	245		502	374	
Lower quartile	248	227		624	560		216	213		383	302	

M: Men W: Women R: Earnings ratio

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 14.2)

Calculate all the earnings ratios (rounded to the nearest one per cent) and

complete the 'R' columns. How do the earnings ratios compare across these occupations?

Statisticians' earnings

Statisticians are in an occupation category that has median weekly earnings of *$*849 for men and *$*707 for women. These are higher than the medians for other occupations in Table 11. Probably the underlying reason, unfortunately, is that management consultants, actuaries and economists are the other groups in the category containing statisticians.

We have now explored enough aspects of gender difference in earnings to conclude that women do still earn less than men, by several different measures, although the overall position is complicated and we have not been able to investigate all the potentially important factors.

In Section 2, we move away from gender comparisons, and look more generally at boxplots.

Exercises on Section 1

Exercise 1 Interpreting a percentile

There are 10 652 000 men in the group represented in Table 6 (in Subsection 1.5). How many of them earned *$*813 or more in that week?

Table 6 is repeated below for convenience.

Percentile	Female	Male
10	253	284
20	296	340
25	316	364
30	335	390
40	380	446
60	497	581
70	574	676
75	619	738
80	671	813
90	820	1083
Median	432	509
Mean	509	635

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 1.2)

Exercise 2 Numerical labelling of a decile boxplot

Figure 10 shows a decile boxplot of the group of female workers in Table 6.

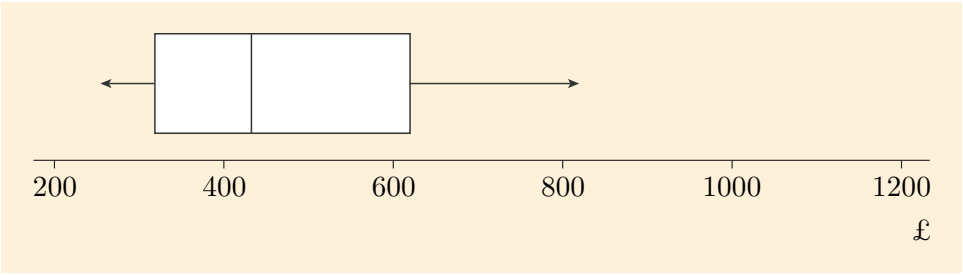


Figure 10 Decile boxplot for female workers

Identify which figures from Table 6 should go by the relevant places on the boxplot to represent the highest and lowest deciles, the quartiles, Q_1 and Q_3, and the median.

Exercise 3 Earnings ratios in the construction industry

As well as providing information on the earnings of people in different occupations, ASHE also gives data on earnings in different industries. Table 12 gives some percentiles of weekly pay, excluding overtime, for male and female workers on adult rates of pay in the construction industry.

Table 12 Weekly pay excluding overtime (rounded to the nearest $) for full-time employee jobs in the construction industry in the UK, 2011

	Female	Male
Highest decile	767	966
Upper quartile	575	689
Median	420	507
Lower quartile	326	401
Lowest decile	277	320

(Source: *Annual Survey of Hours and Earnings*, 2011, Table 4.2)

Calculate the earnings ratios at the deciles, quartiles and median for these data. (Round your answers to the nearest one per cent.) Comment on how the earnings ratios differ across the distribution of earnings.

2 Boxplots and skewness

Subsection 2.1 looks again, briefly, at the skewness of distributions and explores further how skewness is represented and recognised in boxplots. Subsection 2.2 covers some details of drawing boxplots that you have not previously met.

2.1 Recognising skewness

If you look at the boxplots in Figures 8 and 9 (in Subsections 1.6 and 1.8) you can see that, in each case, the right whisker is longer than the left whisker and that the right-hand part of the box is longer than the left-hand part. This pattern is very common in batches of earnings data and, as you know, we describe it by saying that the distribution of earnings data is right-skew.

Let us summarise the ways you have already seen in Units 1 and 2 for recognising skewness. A skew distribution is one which is not symmetric, usually because it has one tail longer than the other. Here is a list of characteristics of a batch of data that can be used to recognise skewness.

- The outline of a stemplot: typical shapes are shown in Figure 11.

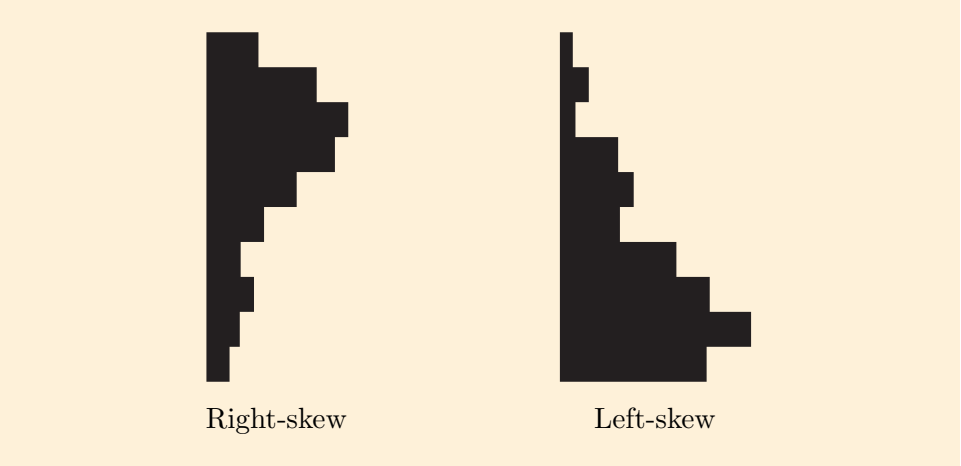


Figure 11 Stemplot shapes

- The whiskers of a boxplot: a long whisker indicates a long tail (though you need to take account of the fact that some of the extreme points might be shown separately on the boxplot).
- The box of the boxplot: in a right-skew batch, the right-hand part of the box (above the median) is longer than the left-hand part of the box (below the median).

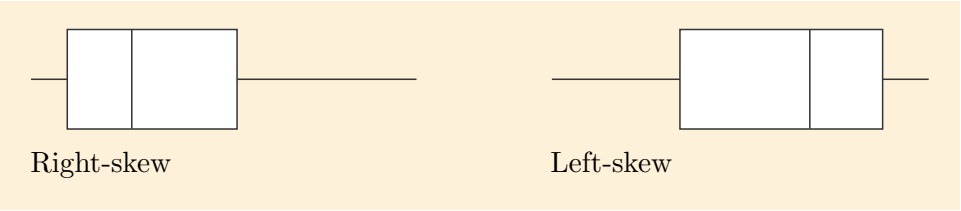


Figure 12 Skewness in boxplots

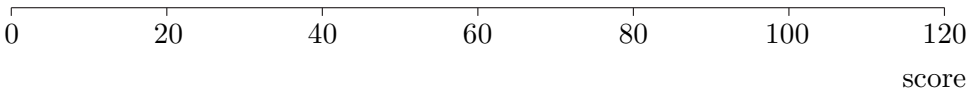
In the following activity, you will get some practice in using the last two of these characteristics to recognise and describe skewness in batches of data. The activity will also serve as a reminder on how boxplots are constructed, as preparation for looking in more detail at the process of drawing boxplots in the next subsection.

Activity 12 Sketching boxplots and recognising skewness

The following five-figure summaries are based on data used in Units 1 and 2. For each batch, sketch its boxplot above the scale provided. (Draw the whiskers going right out to the extremes, for these sketches. For this activity, there is no need to draw them particularly carefully.) Say if a batch is left-skew or right-skew, where appropriate.

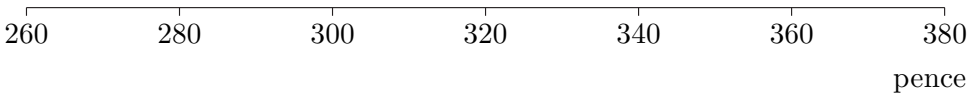
(a) Arithmetic scores data

		79	
$n = 33$	57		88
	7		100



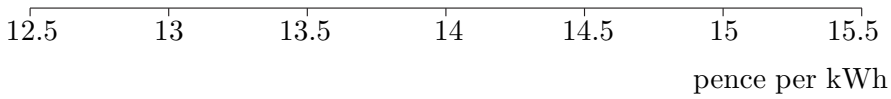
(b) Coffee prices (pence)

		295	
$n = 15$	268		299
	268		369



(c) Electricity prices (pence per kWh)

		13.17	
$n = 15$	12.84		13.83
	12.64		15.03

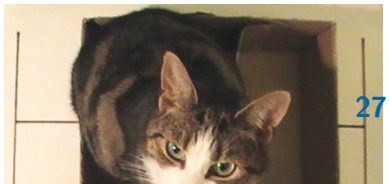


You have now covered the material related to Screencast 3 for Unit 3 (see the M140 website).

2.2 Boxplots: the details

In Subsection 3.3 of Unit 2, you learned about boxplots, and you have since seen several of them in that unit and this one. However, apart from the sketches you were asked to produce in Activity 12, you have not actually been asked to produce any boxplots yourself. You will learn how to do so using Minitab in Section 4. First we give some details, that have not been noted earlier, about the way boxplots are constructed.

You know that the central 'box' of the boxplot shows the positions of the quartiles and the median. But in the boxplots you saw in Unit 2, the whiskers sometimes



went all the way out to the extremes, and sometimes they did not. How is the choice made on how far the whiskers should go? In very broad terms, the lengths of the whiskers depend on how spread out the more extreme values in the batch of data are, compared to the interquartile range of the batch.

Adjacent values

In a boxplot, the whiskers are drawn outwards as far as observations called *adjacent values*.

The **lower adjacent value** is the lowest data value that is within one and a half times the interquartile range of the lower quartile (the lower-end of the box).

The **upper adjacent value** is the highest data value that is within one and a half times the interquartile range of the upper quartile (the upper-end of the box).

To understand exactly how this works, it is easiest to look at an example.

Example 4 Television prices: completing the boxplot

You first met these data in Activity 1 in Unit 2.

Figure 14 shows a stemplot of the prices of small flat-screen televisions. These data came up several times in Unit 2 and you worked with a boxplot of the batch in Subsection 2.1.

	0	9
	1	0
	1	2 3 3 3
	1	4 5 5 5 5
	1	6 6 7
	1	8 8 9
	2	
	2	
	2	4 5
	2	7
$n = 20$	0	9 represents £90

Figure 14 Prices of all flat-screen televisions with a screen size of 24 inches or less on a major UK retailer's website on a day in February 2012

For these data, the median is 150 and the quartiles are $Q_3 = 180$ and $Q_1 = 130$, so that the interquartile range is IQR $= 180 - 130 = 50$.

To find the lower adjacent value, first calculate

$$Q_1 - 1.5 \times \text{IQR} = 130 - 1.5 \times 50 = 130 - 75 = 55.$$

The lowest data value, 90, is greater than this, so the lower adjacent value is 90 and the left-hand whisker on the boxplot extends as far as 90.

Similarly, for the upper adjacent value, first calculate

$$Q_3 + 1.5 \times \text{IQR} = 180 + 1.5 \times 50 = 180 + 75 = 255.$$

The highest data value that does not exceed 255 is 250, so the upper adjacent value is 250, and hence the right-hand whisker on the boxplot extends as far as

250. (Note that it does *not* go all the way to 255, only to the upper adjacent value, the highest value not exceeding 255 in the data. Note also that there is a data value, 270 in this case, that is *above* the upper adjacent value.)

Therefore, the boxplot (not yet complete) with the box and the whiskers looks as in Figure 15.

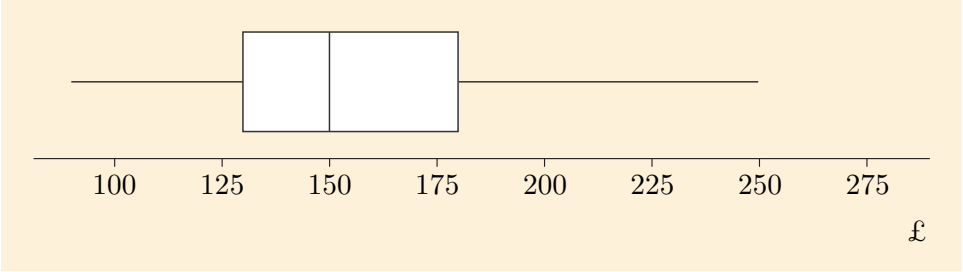


Figure 15 Incomplete boxplot of batch of 20 television prices

The final step is to mark separately any data values that are not covered by the whiskers. In some cases, these may be outliers that do not fit the general pattern of the rest of the batch. In other cases, this is not true, but they are at least *potential* outliers, and the boxplot draws attention to them by plotting them separately.

In these data, there is only one data value not covered by the whiskers, and it is the maximum value, 270. So the resulting boxplot, which you have already seen in Unit 2 (where it was Figure 19 in Subsection 3.3), is shown in Figure 16.

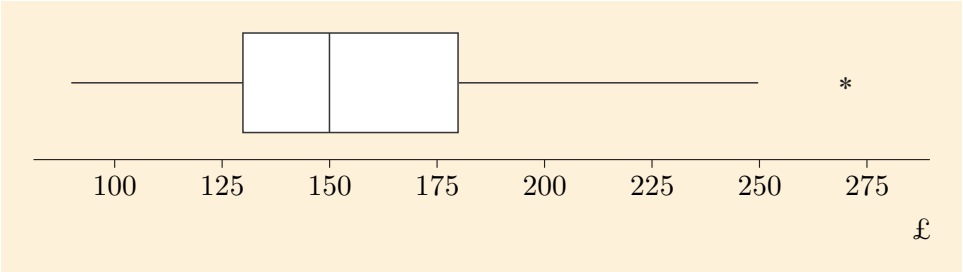


Figure 16 Completed boxplot of batch of 20 television prices

Example 4 is the subject of Screencast 4 for Unit 3 (see the M140 website).

The process of drawing a boxplot can be summarised as follows.

Drawing boxplots

1. The scale for the boxplot must run at least from the minimum to the maximum value in the batch. In M140, the boxplot is drawn so that the scale is horizontal.
2. The 'box' of the boxplot runs from the lower quartile to the upper quartile. Within the box there is a line showing the position of the median.
3. The 'whiskers' of the boxplot are lines, drawn parallel to the scale, that run from the lower quartile to the lower adjacent value, and from the upper quartile to the upper adjacent value. The lower/upper adjacent value is the furthest data value that is within one and a half times the IQR (interquartile range) of the lower/upper quartile.
4. Any individual data values that are not covered by the box or the whiskers are plotted separately (in line with the whiskers). They are potential outliers.

This is the process used in M140, but you must bear in mind that there are no universally accepted rules for drawing boxplots. It is quite common, for instance, to draw boxplots so that they run vertically rather than horizontally. Boxplots always (or almost always!) show the median and the upper and lower quartiles, but the rules defining how long the whiskers extend from the box do vary between different authors and different pieces of software. The approach given here is one of the simplest versions, and is also probably the most common. It is also the approach used by Minitab.

There are many variations on the general boxplot theme – for example you have already come upon decile boxplots, and Figure 17 shows a rather more complicated version of some decile boxplots, taken from the 2011 ONS report on the ASHE results. Here the boxplots run vertically, the boxes are shaded, and different symbols (defined at the side) are used for the median, quartiles and deciles.

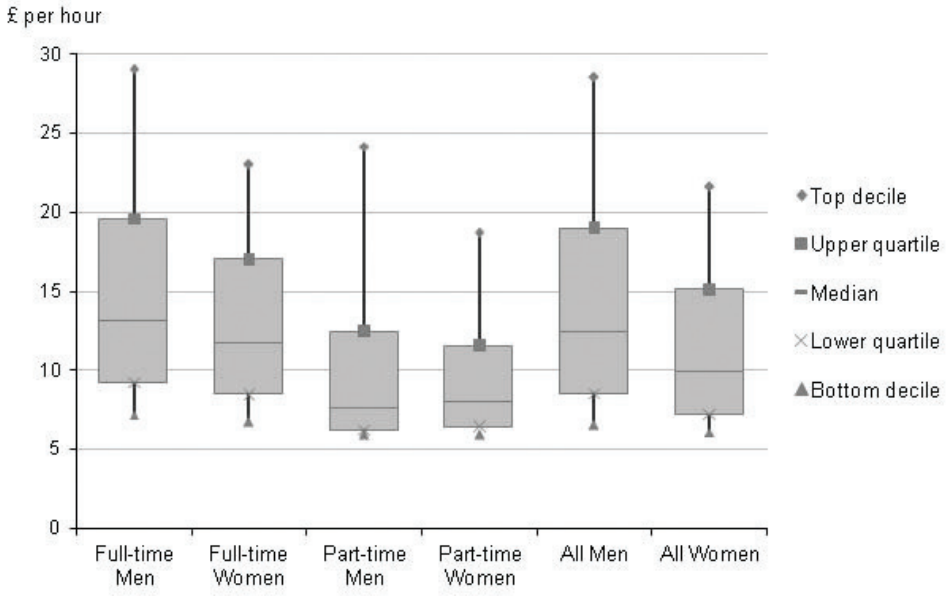


Figure 17 Decile boxplots comparing the hourly earnings of different groups of employees in the UK

You will learn how to use Minitab to produce boxplots in Section 4. But it is worth drawing one or two by hand, to consolidate your understanding of what the different parts mean.

Activity 13 Boxplots of earnings data

The stemplot in Figure 18 gives data on the hourly earnings of 40 female full-time employees in the UK. (These are not real data, but were simulated using a computer on the basis of the distribution of earnings (excluding overtime) for this category from the ASHE 2011 results. However, you should not assume that every feature of this batch of data matches the national figures!)

	0	6 6 6 6 6 6 7 7 7 8 8 8 9 9 9 9 9
	1	0 0 0 0 0 0 1 1 2 2
	1	5 5 5 7 7 7
	2	0 1
	2	9 9
	3	0
	3	
	4	1 3
$n = 40$	0	6 represents £6 per hour

Figure 18 Stemplot of earnings for 40 female employees

(a) Prepare a five-figure summary of the data in Figure 18.

(b) Find the upper and lower adjacent values for these data. Use them, together with the values you calculated for the five-figure summary in (a), to draw a boxplot for the data in Figure 18.

One of the things you looked at in the previous activity was the spread of a batch of data. In the next section, we return to ways of measuring spread.

Exercises on Section 2

Exercise 4 Extracting numbers required for a boxplot from a stemplot

Figure 19 is a stemplot of hourly earnings data for 35 men, generated in a similar way to those for the women in Figure 18.

	0	6 6 7 7 7 7 7 8 8 8 9 9
	1	0 0 1 1 1 2 3 3 4
	1	5 5 5 6 6 7 8 9
	2	1 3 4 4
	2	9
	3	
	3	8
$n = 35$	0	6 represents £6 per hour

Figure 19 Stemplot of earnings for 35 male employees

(a) Find the median and quartiles for these data.

(b) What are the upper and lower adjacent values for the data in Figure 19? Which data values, if any, should be plotted separately on a boxplot?

Exercise 5 A boxplot for petrol consumption data

Figure 20 appeared in Unit 1, and is a stemplot of data on petrol consumption for a particular car.

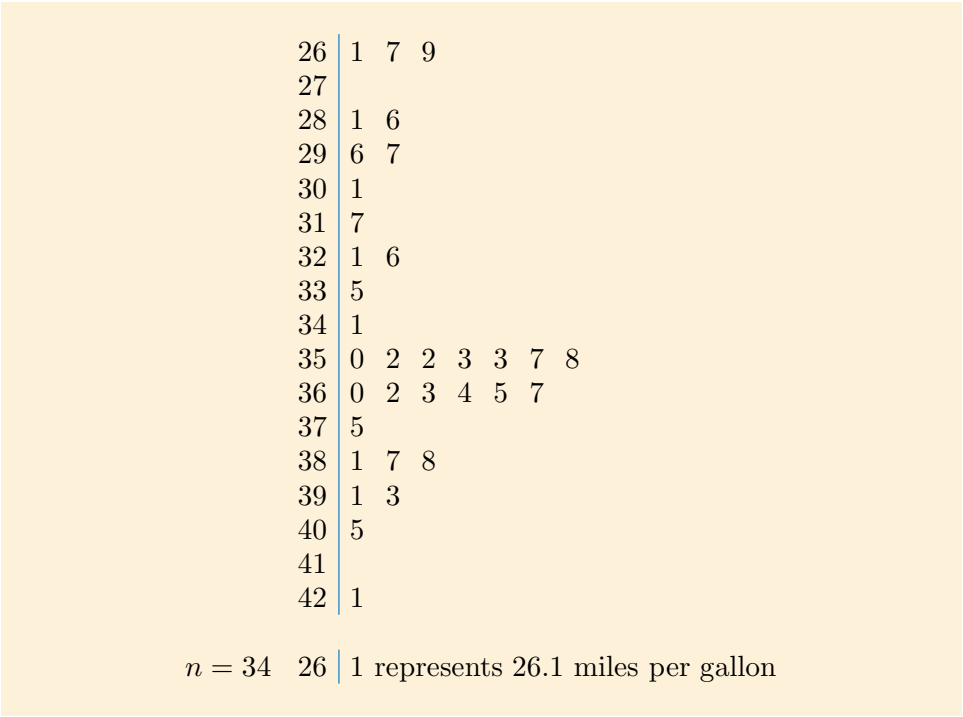


Figure 20 Stemplot of the petrol consumption data from Unit 1

(a) Find the median and quartiles for these data.

(b) What are the upper and lower adjacent values for the data in Figure 20? Which data values, if any, should be plotted separately on a boxplot?

(c) Draw a boxplot of the data in Figure 20. Comment on the skewness of this batch of data.

3 Comparing batches

In this section we return to ways of comparing batches of data. We will do this by considering **summary measures**, which are quantities that summarise aspects of a batch of data, such as its location or its spread. You have already spent some time in previous units working on measures of location, and comparing data in terms of these measures. In Section 2 of this unit, you saw some informal ways of looking at and comparing skewness. In this section, however, we concentrate mostly on the spread of a batch of data and introduce a new measure of spread called the *standard deviation*. This is based on measuring the distances of the data values from the mean of the batch. Subsection 3.2 takes a slight detour, looking at how to calculate the mean and standard deviation for data presented in one form of summary table.

3.1 The standard deviation

This company was first mentioned in Unit 2.

The need to compare batches of data, including looking at their spread, has arisen at our mythical software company, Gradgrind Ltd. The programmers are complaining that they are poorly paid compared to the other staff. Have a look at the data in Table 13 for a few minutes before doing Activity 14. Is their claim justified?

Table 13 Gross weekly earnings ($) of employees of Gradgrind Ltd in April 2011

Programmers	Others
465	346
484	376
620	391
654	391
855	415
858	465
	830
	843
	876
	1627

We can begin investigating this question by looking at summary measures of locations and spread that you already know how to calculate.

Activity 14 Calculating some summary measures

Treating the programmers as one batch and 'others' as a second batch, for each of these batches find the following:

(a) the mean (b) the median (c) the interquartile range.

In Activity 14, you calculated the following two different measures of overall location.

- The mean: the two batches have the same mean ($656).
- The median: the median earnings of the programmers ($637) is higher than that of the other staff ($440).

In both cases, the median is less than the mean, which, as you know from earlier in the unit, typically happens in right-skew batches.

You have also calculated the following measure of spread.

- The interquartile range: the other staff earnings have a slightly higher IQR ($464) than that of the programmers' earnings ($377).

Now we shall introduce another measure of spread: the standard deviation. This measure of spread is related to the mean in the same way that the interquartile range is related to the median – by the method of calculation. Both the median and the IQR are found by ordering the data and finding the values at a particular position (or between two particular positions) of the ordered list of values. The standard deviation, on the other hand, is found by doing calculations similar to those used in calculating the mean: finding sums and dividing by the batch size (or something similar to the batch size).

The basic idea is to calculate a numerical measure that reflects how much the data values spread out or, more precisely, how much the x values spread away from the 'centre' of the batch. We shall represent the 'centre' of a batch by the mean, $\overline{x}$

The deviation

For each data value in a batch we have a **deviation of the data value from the mean**, or just **deviation** for short. If we have a data value x, and the batch mean is $\overline{x}$, then the deviation is $x - \overline{x}$.

As a simple example, suppose a batch consists of just five data: 6, 7, 2, 6 and 4. Then the mean of the batch equals

$$\frac{6+7+2+6+4}{5} = 5.$$

Taking 5 from each data value, the first deviation is $6-5=1$ and the others are $2, -3, 1$ and -1.

The size of the deviation

The deviations measure how far the data values in a batch are from the batch mean.

- If a data value is exactly equal to the mean, then the deviation will be zero.
- If a data value is close to the mean, then the deviation will be a small number, near zero.
- If a data value is a long way above the mean, then the deviation will be large and positive.
- If a data value is a long way below the mean, then the deviation will be large and negative.

Thus if a batch has a large spread, its data values will tend to be a relatively long way away from the mean, so the deviations will tend to be large in size. (Large negative numbers and large positive numbers are both large in size.)

The *standard* deviation, as you will see, is just a kind of average deviation – so it will be larger in batches that are more spread out. However, there is an awkward aspect – the deviations have signs, that is, some are positive and some are negative. If all we are interested in is how far the data values are from the batch mean, then the sign is not very interesting. A value 10 units *below* the mean is just as far away from the mean as a value 10 units *above* the mean, so in looking at the spread of the batch, these values should be treated in the same way even though one has a deviation of -10 and the other of 10.

There is more than one way to get round this sign issue, but the one used in calculating the standard deviation is to *square* the deviations. Think again of two data values that have deviations -10 and 10. If these deviations are squared, the squared deviations are both 100 (because the square of a negative number is positive). So squaring them means that the issue of the signs is dealt with.

Activity 15 Calculating the sum of squared deviations

Table 14 shows the data values for the programmers, as given originally in Table 13. The first column is completed, and gives all the data values. The mean of these values, $\overline{x}$, is 656. The second column is also completed, and gives this batch mean (which is the same for the whole batch, of course). The entries in the other two columns are not complete. The third column gives the deviation corresponding to each data value, and the deviation for the first data point is filled in as -191. This was calculated from $x - \overline{x} = 465 - 656 = -191$. The fourth column gives the squared deviations. It is again completed only for the first data point, with $(-191)^2 = 36\,481$.

Table 14 Sum of squared deviations

Data x	Mean $\overline{x}$	Deviation $(x-\overline{x})$	Squared deviation $(x-\overline{x})^2$
465	656	-191	36 481
484	656		
620	656		
654	656		
855	656		
858	656		
	$\sum$		

Complete the table by calculating the remaining values for the third and fourth columns, followed by the sum of each of these columns.

In Activity 15, the sum of the deviations was zero. This also happened with our first simple example, where the data values were 6, 7, 2, 6, 4 with deviations 1, 2, -3, 1, -1 (and $1+2-3+1-1=0$). In fact this will always happen, because there will always be some values below the mean (negative deviations) and some above (positive deviations), and because of the way the mean is calculated: the positives and negatives will always exactly cancel out, giving a zero sum. (That is another reason why we cannot just use the average deviation as a measure of spread.)

However, the sum of the *squared* deviations is not zero, and we could compare batches in terms of spread by looking at this sum – or better, by looking at the *mean* of the squared deviations in a batch. However, there is a small extra complication here. For the programmers, there are six squared deviations, so you might expect to calculate the mean by dividing the sum of the squared deviations by the batch size, 6. But it turns out that dividing by one less than the batch size, that is, by 5 in this case, gives a measure of spread with better statistical properties. (You may have come upon definitions in other places where one is not subtracted in the divisor – and many calculators will calculate the standard deviation in both ways, dividing by either the batch size n or by $n-1$, so you should check that you know how to make you calculator use the version that divides by $n-1$ before using it to find a standard deviation in M140.)

The variance

The quantity obtained, for a batch of size n, by calculating the sum of squared deviations and dividing by $n-1$, is called the **variance** of the batch. It is a measure of spread.

For the programmers, the sum of squared deviations is 147 770, so the variance is

$$\frac{147\,770}{5} = 29\,554.$$

However, this calculation has one drawback. With these data on earnings, the unit of measurement of the original data is *$*. The batch mean is also in *$*, so the deviations are measured in *$* too. But the *squared* deviations have units of '$\$^2$', which look rather strange and are not the original units of measurement of the data. So the final stage in calculating the standard deviation is to take the square root of the variance. That changes the units of measurement back to *$*. In our

example, then, for the programmers the standard deviation is $\sqrt{29\,554} = 171.912\,77$. As with the mean, we shall generally round the standard deviation to one decimal place more than the original data, so in this case to $171.9.

We shall generally denote the standard deviation of a batch of data by the letter s. The steps to calculate its value for a batch of data are summarised in the following box. (This is called 'Method 1' because there is also a 'Method 2' that you will meet later in this subsection.)

Calculating the standard deviation: Method 1

.1 Calculate the mean $\overline{x} = \dfrac{\text{sum}}{\text{size}} = \dfrac{\sum x}{n}$.

.2 Calculate the deviations $(x - \overline{x})$.

.3 Square the deviations to give $(x - \overline{x})^2$.

.4 Calculate the variance by summing the squared deviations, to give $\sum(x - \overline{x})^2$, and dividing by $(n - 1)$: that is,

$$\text{variance } (s^2) = \frac{\sum(x - \overline{x})^2}{n - 1}.$$

.5 Calculate the standard deviation as $s = \sqrt{\text{variance}}$.

Check that you can do these calculations by finding the standard deviation of the earnings for Gradgrind's other workers in the following activity.

Activity 16 Calculating a standard deviation using Method 1

Table 15 shows the data values for the other workers, as given originally in Table 13. The first two columns are completed. Calculate the values for the third and fourth columns, and the sum of the fourth column. Then calculate the variance and standard deviation for this batch.

Table 15 Sum of squared deviations

Data x	Mean $\overline{x}$	Deviation $(x - \overline{x})$	Squared deviation $(x - \overline{x})^2$
346	656		
376	656		
391	656		
391	656		
415	656		
465	656		
830	656		
843	656		
876	656		
1627	656		
	$\sum$	0	

You may well be thinking that it is quite a tedious business to calculate a standard deviation, using the kind of methods that have been described, using a calculator – and so far the biggest batch you have looked at had only 10 values

in it. In practice you will not be expected to do this kind of calculation too often. In M140, and indeed in the real world, standard deviations would typically be calculated in a computer program (Minitab for M140), or perhaps (for small batches) using some special features on a calculator.

If your calculator has features that allow easier calculation of standard deviations, you should use them where appropriate. However, because different calculators deal with this sort of thing differently, it is important that you check how to do so on your own calculator. Also, as previously mentioned, many calculators give you the choice between dividing by n and dividing by $n-1$ in calculating the variance, and you must therefore be sure to use the $n-1$ version for M140.

However, there is an alternative method of calculating standard deviations that can be useful on calculators. This second method is illustrated again using the data for Gradgrind's programmers.

The difference between Method 1 (the method you have used so far) and Method 2 (the new method) lies in how the quantity $\sum(x-\overline{x})^2$ is calculated. Up to now, $\sum(x-\overline{x})^2$ has been found by first finding the deviations by subtracting the mean from each data value, then squaring the deviations, and then adding them up. But it turns out that

$$\sum(x-\overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}.$$

This can be proved mathematically, but we shall not do that here as you do not need the proof for M140. We shall just illustrate that it works for the Gradgrind programmers' earnings.

Table 16 shows again, in the first column, the data values for the programmers from Table 13. The second column contains the squares of the data values – not the squares of the deviations (i.e. the squared deviations) that you saw before, but the squares of the original data values. (So, for example, the first value in the second column is $465^2 = 216\,225$.) At the bottom of the table, the sums of both columns are given.

Table 16 Squared data for programmers

	Data x	Squared data x^2
	465	216 225
	484	234 256
	620	384 400
	654	427 716
	855	731 025
	858	736 164
$\sum$	3936	2 729 786

Method 2 requires us to calculate $\sum(x-\overline{x})^2$ as $\sum x^2 - \frac{(\sum x)^2}{n}$. Now, $\sum x^2$ is the sum of the second column in Table 16, the sum of the squared data values, 2 729 786. Note that $\sum x^2 = \sum(x^2) = (\sum x^2)$. However, the quantity $(\sum x)^2$ is *not* the same as $\sum x^2$. It is the square of the sum of the data values (not the sum of the squares) and is found by adding the values in the first column and

then squaring their sum. That is, $(\sum x)^2 = (3936)^2$. Therefore,

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &= 2\,729\,786 - \frac{3936^2}{6}\\ &= 2\,729\,786 - \frac{15\,492\,096}{6}\\ &= 2\,729\,786 - 2\,582\,016 = 147\,770.\end{aligned}$$

This result is exactly what you found, using Method 1, for the sum of the squared deviations, $\sum(x-\overline{x})^2$, in Activity 15. The rest of this calculation of the standard deviation follows exactly as it did in Method 1. The variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{147\,770}{5} = 29\,554.$$

The standard deviation is

$$s = \sqrt{\text{variance}} = \sqrt{29\,554} = 171.9 \text{ (to one decimal place)}.$$

You have now covered the material related to Screencast 5 for Unit 3 (see the M140 website).

The steps for calculating the standard deviation of a batch of data by Method 2 are as shown in the following box.

Calculating the standard deviation: Method 2

1. Calculate the sum of the data values, $\sum x$.
2. Calculate the sum of the squares of the data values, $\sum x^2$.
3. Calculate the sum of the squares of the deviations, $\sum(x-\overline{x})^2$, as
$$\sum x^2 - \frac{(\sum x)^2}{n}.$$
4. Calculate the variance by dividing $\sum(x-\overline{x})^2$ by $(n-1)$: that is,
$$\text{variance } (s^2) = \frac{\sum(x-\overline{x})^2}{n-1}.$$
5. Calculate the standard deviation as $s = \sqrt{\text{variance}}$.

So there are two methods of calculating the standard deviation. Both methods give the same answer, although if you are using a basic calculator it is generally easier to use Method 2. The methods have already been described, but a briefer way to describe them is in the following box.

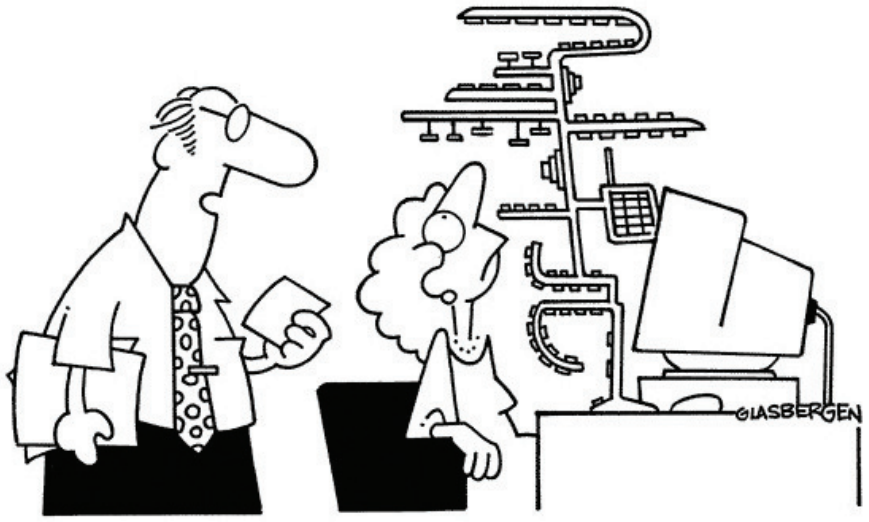
Sum of squared deviations

- Method 1: Calculate the mean $\overline{x}$, subtract it from each data value and hence work out $\sum(x-\overline{x})^2$ directly.
- Method 2: Calculate the sum of the data values, $\sum x$, and the sum of the squares of the data values, $\sum x^2$, and hence work out
$$\sum(x-\overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}.$$

Then, by either method,

Variance: $s^2 = \dfrac{\sum(x-\overline{x})^2}{n-1}$,

Standard deviation: $s = \sqrt{\text{variance}}$.



'It's the new keyboard for the statistics lab. Once you learn how to use it, it will make computation of the standard deviation easier.'

Activity 17 Calculating a standard deviation using Method 2

Using Method 2, calculate the standard deviation for Gradgrind's ten other staff using the data in the first column of Table 17.

Table 17 Squared data for 'other staff'

Data x	Squared data x^2
346	
376	
391	
391	
415	
465	
830	
843	
876	
1627	
$\sum$	

Start by completing the second column of the table with the squares of the data values and calculating the column sums.

Then calculate the sum of the squared deviations using

$$\sum (x - \overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}.$$

Finally, calculate the variance and standard deviation. (You should, of course, get the same variance and standard deviation as in Activity 16.)

To summarise our findings on spread for the two batches in Table 13, the interquartile ranges (IQR) and the standard deviations (s) are given below.

	Programmers	Other staff
IQR	377	464
s	171.9	403.6

Both measures of spread show that the data for the other staff is more widely spread out than the data for the programmers. But the difference looks much greater in terms of s, where the standard deviation for the other staff is considerably more than twice as large as that for the programmers. The reason is essentially that the batch of earnings for the other staff contains one figure, *$*1627, that is considerably bigger than all the others in that batch. Just as the mean is not a resistant measure and can be strongly affected by one or two values near the extremes, the standard deviation is also not a resistant measure of spread. The interquartile range, on the other hand, *is* a resistant measure of spread. (This difference between the two measures will be mentioned in Subsection 3.3.)

These ideas were introduced in Subsection 1.4 of Unit 2.

Activity 18 Standard deviation for all employees

Using Method 2, calculate the mean and the standard deviation of the combined batch consisting of the earnings of all the employees of Gradgrind Ltd in Table 13. (Hint: you can work out the necessary sums without having to start from scratch, by using the calculations in Table 16 and in the solution to Activity 17.)

You have now covered the material needed for Subsection 3.1 of the Computer Book.

3.2 Calculating the mean and standard deviation for grouped data

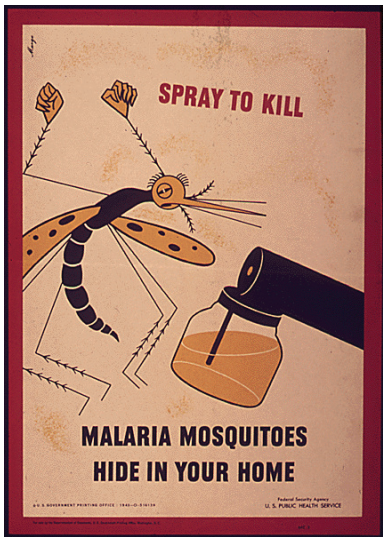


Figure 21 A U.S. government poster from the 1940s aimed at combatting malaria

Suppose we sampled a large number of families and noted the number of children in each. We could record this information as: 'The first family had two children, the second family had none, the third family had . . . '. However, for a large sample the list would be long. It would be easier to group the families together according to the number of children and just record the number of families that had no children, the number with one child, the number with two children, and so on. If no family had more than 15 children then this list would have 16 or fewer items (0 to 15 children). Data recorded in this form are called **grouped data**.

When a batch contains a large number of items, calculating the batch mean and standard deviation might seem a daunting task. With grouped data, however, the amount of labour that is involved depends on the number of groups, and not on the number of individual items. This will be illustrated using the following data, consisting of 50 000 individual items but only five groups.

The data in Table 18 come from a study of the way in which the malaria parasite *Plasmodium falciparum* invades red blood cells. A sample of blood was taken from a patient with malaria caused by this species of parasite, and the number of parasites in each of 50 000 red blood cells in the sample was counted.

Table 18 Malaria parasites in red blood cells

Number of parasites per cell x	Frequency f
0	40 000
1	8 621
2	1 259
3	99
4	21

(Source: Wang, C.C. (1970) 'Multiple invasion of erythrocyte by malaria parasites', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 64, issue 2, pp. 268–270)

The data presented in Table 18 is an example of grouped data. Each row of the table corresponds to a different group. The numbers in the table mean that there were (exactly!) 40 000 cells that contained no parasites at all, 8621 with exactly one parasite, 1259 with two parasites, and so on. No cell contained more than four parasites.

The term **frequencies** is used for numbers (such as those in the second column of Table 18) which tell us how frequent the corresponding data values are. We shall generally denote frequencies by f.

How do we calculate the mean and standard deviation of the grouped data in Table 18? It turns out that we need the batch size, the sum of the values and the sum of their squares. The following example shows how these can be obtained.

Example 5 Malaria parasites: calculations with grouped data

Obtaining the batch size for the grouped data in Table 18 is easy. There are 40 000 blood cells that had no parasites, 8621 had one, 1259 had two, 99 had three, and 21 had four, so we obtain the batch size by adding the numbers in the frequency column. That is, the batch size equals the total number of blood cells:

$$\sum f = 40\,000 + 8621 + 1259 + 99 + 21 = 50\,000.$$

Now, for the sum of the data values, imagine that for all of the 50 000 cells we decided to write out in a row the number of parasites in each cell and add them up. In this sum, there would be 40 000 zeros, which contribute $0 \times 40\,000$, that is, zero, to the sum; 8621 ones, which contribute $1 \times 8621 = 8621$ to the sum; 1259 twos which contribute $2 \times 1259 = 2518$ to the sum, and so on. That is, to get the sum of the data values, we calculate, for each row of the table, the product xf, and then we sum these products. This is shown in Table 19.

Table 19 Malaria parasites: calculating the sum

x	f	xf
0	40 000	0
1	8 621	8 621
2	1 259	2 518
3	99	297
4	21	84
$\sum$	50 000	11 520

That is, the total number of parasites in the whole batch of blood cells can be obtained by summing the number of parasites in each group:

$$\sum xf = 0 + 8621 + 2518 + 297 + 84 = 11520.$$

The sum of the squares of the data values can be calculated in a similar way. The 50 000 zeros each contribute $0^2 = 0$ to the sum, so their total contribution to the sum of squares is 0. The 8621 ones each contribute $1^2 = 1$, so their total contribution to the sum of squares is $8621 \times 1^2 = 8621$. The 1259 twos each contribute $2^2 = 4$, so their total contribution to the sum of squares is $1259 \times 4 = 5036$. And so on. That is, we get the sum of squares by multiplying x^2, the *square* of each data value, by the corresponding frequency, f, and summing the resulting quantities – that is, we calculate $\sum x^2 f$. These calculations are shown in Table 20.

Table 20 Malaria parasites: calculating the sum of the squares

x	x^2	f	xf	$x^2 f$
0	0	40 000	0	0
1	1	8621	8621	8621
2	4	1259	2518	5036
3	9	99	297	891
4	16	21	84	336
		$\sum f = 50\,000$	$\sum xf = 11\,520$	$\sum x^2 f = 14\,884$

So, the sum of the squares of the data values is

$$\sum x^2 f = 14\,884.$$

Now we have calculated the batch size, the sum of the values and the sum of their squares, we can calculate the mean and standard deviation for the grouped data using the following method. It is the same as Method 2 for ungrouped data, but with n, $\sum x$ and $\sum x^2$ replaced with $\sum f$, $\sum xf$ and $\sum x^2 f$.

Mean and standard deviation from grouped data

Denoting the data values by x and the corresponding frequencies by f:

1. Construct a table similar to Table 20 to calculate the batch size, $n = \sum f$, the sum of the data values, $\sum xf$, and the sum of the squares of the data values, $\sum x^2 f$.
2. Calculate the mean as $\overline{x} = \dfrac{\sum xf}{n}$.
3. Calculate the sum of the squares of the deviations as
$$\sum (x - \overline{x})^2 f = \sum x^2 f - \frac{(\sum xf)^2}{n}.$$
4. Divide the result of step 3 by $n - 1$, giving
$$\text{variance } (s^2) = \frac{\sum (x - \overline{x})^2 f}{n - 1}.$$
5. Calculate the standard deviation as $s = \sqrt{\text{variance}}$.

Example 6 Malaria parasites: calculating the mean

Using the totals in Table 20, the batch size of the malaria parasite data is

$$n = \sum f = 50\,000$$

and the sum of the data values is

$$\sum xf = 11\,520.$$

So we can calculate the mean as:

$$\overline{x} = \frac{\text{total number of parasites}}{\text{total number of blood cells}} = \frac{\sum xf}{\sum f}$$
$$= \frac{11\,520}{50\,000} = 0.2304 \simeq 0.23.$$

Therefore, on average, there are approximately 0.23 parasites per red blood cell.

Activity 19 Malaria parasites: calculating the standard deviation

Using the totals given in Table 20, calculate the standard deviation for the number of parasites in a red blood cell.

Activity 20 Flying bomb hits

The (grouped) data in Table 21 come from a study of the sites where flying bombs hit South London during World War Two (see Figure 22). The whole area was divided into 576 squares, each one quarter of a square kilometre, and the number of hits in each square was counted. (The purpose of the study was to investigate whether the flying bombs were precisely aimed, or instead just aimed in the general direction of London where they then landed at random. The conclusion was that the bombs were *not* precisely aimed.)

Calculate the mean number of hits per square, and the standard deviation of the number of hits per square.

Table 21 Flying bomb hits

Number of hits per square x	Frequency f
0	229
1	211
2	93
3	35
4	7
5	1

(Source: Clarke, R.D. (1946) 'An application of the Poisson distribution', *Journal of the Institute of Actuaries*, vol. 72, p. 481)

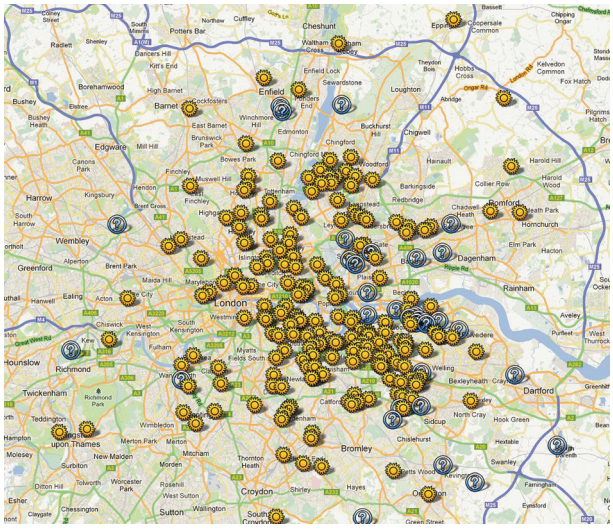


Figure 22 Map of London indicating where flying bombs landed

You have now covered the material related to Screencast 6 for Unit 3 (see the M140 website).

3.3 Deciding which measure to use

You know how to find the two major measures of spread: the standard deviation and the interquartile range. So you may well be wondering how to decide which one to use. This is very similar to deciding whether to use the mean or the median as a measure of location (see Subsection 1.4 of Unit 2). Among the factors to bear in mind are the following.

- *Consistency*. Owing to their similarities, these measures tend to be used in the following pairs:
  - the mean and standard deviation
  - the median and interquartile range.

  It is quite rare to present an analysis which uses the median to measure location and the standard deviation to measure spread, or the mean to measure location and the interquartile range to measure spread. The reason that these measures tend to go together in this way is to do with the underlying arithmetic operations that are used in their calculation. Essentially, the mean and standard deviation involve adding, while the median and interquartile range are based on counting.
- *Purpose*. For the next stage in your analysis of the data, you may require one measure rather than the other. Later in M140 you will meet some methods that use medians, and others that use means and standard deviations. When choosing your measure, you should bear in mind which other methods you are going to use when analysing the data.
- *Resistance*. Just as the mean is less resistant than the median, the standard deviation is less resistant than the interquartile range. The greater resistance of the interquartile range was referred to in Subsection 3.1 (just before Activity 18), in the context of the Gradgrind earnings figures. It is explored further in the following activities.

A *resistant* measure is one which is not particularly affected by changes in the values near the extremes. (See Subsection 1.4 of Unit 2.)

Activity 21 Resistance of measures of spread

Suppose that the lowest and highest values in the batch of earnings of Gradgrind's 'other' staff (Table 13 of Subsection 3.1) were altered to 246 and 1727. Find the interquartile range and the standard deviation of the new batch. The *original* data for the other staff is repeated here for convenience.

346 376 391 391 415 465 465 830 843 876 1627

There are two key points to emerge from the calculations for Activity 21:

- Despite having altered two of the values, the interquartile range remains unchanged from its original value of 464.
- By making the distribution more spread out, the two alterations have increased the standard deviation from its original value of 403.6 to 440.0. (The value of 403.6 for the standard deviation was calculated in Activity 16 of Subsection 3.1.)

In the next section, you will see how to use Minitab to calculate some of the summary measures from this unit and Unit 2, and how to draw boxplots.

Exercises on Section 3

Exercise 6 Winter energy consumption

Calculate the mean and standard deviation of each of the two batches of data in Table 22. (MWh is the usual abbreviation for megawatt hours.)

Table 22 Winter energy consumptions (MWh) of ten houses in Bristol before and after insulation

Before insulation	After insulation
12.1	12.0
11.0	10.6
14.1	13.4
13.8	11.2
15.5	15.3
12.2	13.6
12.8	12.6
9.9	8.8
10.8	9.6
12.7	12.4

Exercise 7 Counting Macaulay's words

In a study aimed at developing a method of characterising an author's style, samples of ten words were taken from the beginning of each of 100 randomly chosen lines from the printed text of *Macaulay's Essay on Milton* (T.B. Macaulay, 1895). In each ten-word sample, the number of times that the three article words 'the', 'a' and 'an' appeared was counted. The data appear (in grouped form) in Table 23. Calculate the mean and standard deviation for these data.

Table 23 Articles in Macaulay

Number of articles x	Frequency f
0	27
1	44
2	26
3	3

(Source: Bailey, B.J.R. (1990) 'A model for function word counts', *Applied Statistics*, vol. 39, no. 1, pp. 107–114)

4 Computer work: summary measures and boxplots

In Unit 1 you were introduced to the Minitab software. In this section, you will learn how to use Minitab to calculate some of the summary measures you have learned about in Units 2 and 3 – means, medians, quartiles, interquartile range and standard deviations. You will also learn how to draw boxplots with Minitab.

You should now turn to the Computer Book and work through Subsection 3.1, if you have not already done so, followed by the rest of Chapter 3.

5 Prices and earnings

A central question which ran through Unit 2 was: *Are people getting better or worse off?* You saw there that this is a difficult question to answer precisely. For a start, there are many different factors to take into account; also, what may be true for one person will not necessarily be true for another. However, it is clear that two key factors, prices and earnings, are highly relevant to the question. Unit 2 and this unit have been devoted to considering how these two factors can be measured.

In this final section, we shall attempt to bring together prices and earnings to see how they compare over time. We start with earnings and consider a drawback inherent in data taken from the Annual Survey of Hours and Earnings (ASHE).

You know already from Subsection 1.3 that the ASHE does not cover the earnings of self-employed people (or indeed unemployed people). Another major drawback, for some purposes, is that the survey is carried out only once a year. This is far less frequent than the price indices (RPI and CPI) that you met in Unit 2, which are published monthly. In monitoring the economy, data that come out only once a year are not frequent enough for the government. Further data on earnings are available from the Labour Force Survey, which is a major data source for (among other things) measures of unemployment that go beyond simply counting up how many people are claiming out-of-work benefits; it also collects data on earnings from its respondents. (The Labour Force Survey collects its data from individual respondents, not their employers. This leads to some issues of accuracy because, in some cases, the responses are given by a person other than the actual income earner if they are not available for interview, and that person may not know the level of income closely enough.)

The Labour Force Survey publishes earnings data every three months, but even that is not frequent enough for some purposes. We shall therefore turn to an alternative measure of income, published monthly, called the Average Weekly Earnings (AWE) index.

5.1 The Average Weekly Earnings (AWE) index

In this subsection, we are going to look at an index which measures *changes* in most people's main source of income: their earnings. This index is the Average Weekly Earnings (AWE) index, calculated by the ONS once a month. We will describe briefly how the data used for the AWE index are collected and how the index is calculated. Then, at the end of the section, we will describe how the AWE index, together with the CPI can be used to investigate the economic health of society.

The AWE index is based on the Average Weekly Earnings measure. Here are some extracts from the official description of the measure.

> The Average Weekly Earnings (AWE) measure is the Office for National Statistics' (ONS's) lead indicator of short-term changes in earnings. ... AWE is published monthly and is designed to capture changes in average earnings of employees in Great Britain. ... Average Weekly Earnings is calculated from returns to the Monthly Wages and Salaries Survey (MWSS), and is weighted to be representative of the Great Britain economy as a whole. The self-employed, HM Armed Forces and Government Supported Trainees are excluded from the statistics.
>
> (Source: ONS (2011) *Quality and Methodology Information for Average Weekly Earnings*)

To find out what is meant here by earnings, we must look at the data which are used to calculate the AWE measure. These data come from a survey called the Monthly Wages and Salaries Survey. Each month, a sample of around 9000 firms in Great Britain receives a simple questionnaire whose main purpose is to obtain the following figures.

The survey does not cover Northern Ireland.

- The number of monthly-paid employees receiving pay in that month.
- The total gross amount paid to all monthly-paid employees during that month.
- The number of weekly-paid employees receiving pay during the last pay-week in that month.
- The total gross amount paid to all weekly-paid employees during that week.

Employers are specifically asked to include overtime and holiday pay, as well as other additional payments, in their returns, and not to make deductions for tax, national insurance and pension contributions, etc. They are asked to exclude fees paid to directors, trainees on government schemes, and certain other categories. The questionnaire also asks employers to provide data on arrears of pay, such as those arising from backdated pay increases. These arrears are included in the gross pay amounts, but are subtracted from them when calculating the index based on the AWE measure.

So we now know that, in the Average Weekly Earnings measure, **earnings** simply means *the gross amounts paid to employees (excluding pay arrears)*. Now let us look at how these data are used to calculate an index that measures *changes* in the overall level of earnings.

1. The first step is to find the average weekly earnings for each of the businesses in the sample. For those paid monthly, their pay is converted to a weekly figure. Then the average weekly earnings is calculated by dividing the total gross amount paid to all employees by the number of employees.
2. The next step is to calculate a national figure for average weekly earnings, as a weighted average of the data for each business. The weighting that is used

is rather complicated, taking into account the size of the business, the industry it is in, and whether it is public or private, but in principle this is simply a weighted average (that is, a weighted mean) like those you met in Unit 2.

3. Next, the average weekly earning figures are *seasonally adjusted.* This means that they are adjusted to allow for the effect of changes in earnings levels that occur regularly at fixed times of the year. Thus, if the AWE measure shows an increase between one month and the next, this means that wages have gone up (on average) that month by *more* than they would normally be expected to increase at that time of year.
4. Finally, the Average Weekly Earnings index is calculated by comparing the national average weekly earnings with the corresponding figure for the base year. At the time of writing (2012), the base year is 2000 and the AWE index for 2000 is set at 100. Like the CPI (but unlike the RPI), for the AWE index the average earnings are compared with the average over the whole base year, and not just a single month. That is, the value of the AWE index for a given month is

$$100 \times \frac{\text{average weekly earnings in that month}}{\text{average weekly earnings in the base year}}.$$

This differs from both the RPI and CPI, in that the comparison is made *directly* with base date. With the AWE index, there is no chaining.

The Average Weekly Earnings index thus provides information on changes in the overall level of earnings in Great Britain. Figures for the index, together with the average weekly earnings on which it is based, are published each month on the ONS website. Although versions of the index are published for different sectors of the economy, we shall only use the index for the whole economy.

You have now learnt about two sources of earnings data: the AWE index and the ASHE. The most important difference (apart from their frequency) is that, whilst ASHE measures the *level* of earnings of different groups, the AWE index is primarily designed to measure *changes* in earnings. (Data on average earnings levels *are* published for AWE, but its design is primarily aimed at looking at changes.) Another difference is that, whereas ASHE covers a sample of *individuals*, the AWE index covers a sample of *firms*. A simple way of comparing the two sources would be to say that the AWE index uses a quick, relatively cheap survey which is designed for a specific purpose (looking at overall changes in earnings levels), whereas ASHE is a more detailed, general-purpose survey.

No doubt you will have noticed a considerable degree of similarity between the calculation of the AWE index and that of the RPI and CPI described in Unit 2. We are now ready to put together the AWE index and one of the price measures, the CPI, and compare price and income changes over the period from 2001 to 2011.

5.2 Comparing the AWE with the CPI

> 'Annual income twenty pounds, annual expenditure nineteen and six, result happiness. Annual income twenty pounds, annual expenditure twenty pounds ought and six, result misery.'
>
> (Source: Mr Micawber in Charles Dickens' *David Copperfield* (1850))

Our motivation for investigating the AWE index is to use it in conjunction with a *price* index to measure the economic well-being of the nation. For this purpose we shall use the CPI rather than the RPI.

The balance between pay and prices is often a precarious one and Mr Micawber's comment (quoted above) is particularly pertinent when pay and

prices are rising rapidly, for it is then often difficult to ensure that both rise in step. A statistical Mr Micawber might well say: 'AWE index exceeds CPI, result happiness. CPI exceeds AWE index, result misery'. Because the CPI measures only price changes, to make such a comparison we must compare changes in both the indices over a given period. In Unit 2 you learned how to express changes in the RPI and CPI. We shall now see how to do the same thing with the AWE index.

Figure 23 An image of Mr Micawber

Example 7 Calculating a change in the AWE index

The value of the AWE index for January 2011 was 144.7 and the value for January 2012 was 145.4. Therefore its value in January 2012 as a percentage of its January 2011 value was

$$\frac{145.4}{144.7} \times 100\% = 100.483\,76\% \simeq 100.5\%.$$

Thus its increase over the year January 2011 to January 2012 was 0.5% of its January 2011 value. That is, earnings on average did increase over that year, but only by half of one per cent.

Figure 24 Another CPI. Does it fill you with AWE?

Activity 22 Percentage change in the AWE index

The value of the AWE index for January 2000 was 98.2 and the value for January 2001 was 103.6. By what percentage of the January 2000 value did the AWE index increase over the year?

To compare changes in the CPI with those in the AWE index we need to calculate the changes in both indices. This is shown in Figure 25, covering the period from 2001 to 2011. The graph reveals that, during the middle years of the decade, the increase in the AWE index remained fairly steady at around 4% per annum and was at a level consistently higher than the average price rises (shown by the CPI). After the beginning of 2007, though, earnings increases fell below the level of price increases, and generally remained so right up to the end of the period covered. (The extreme negative values of the AWE index change in early 2008, and the high positive values in early 2009 and 2010, are due to large fluctuations in bonus pay, particularly in the financial sector.)

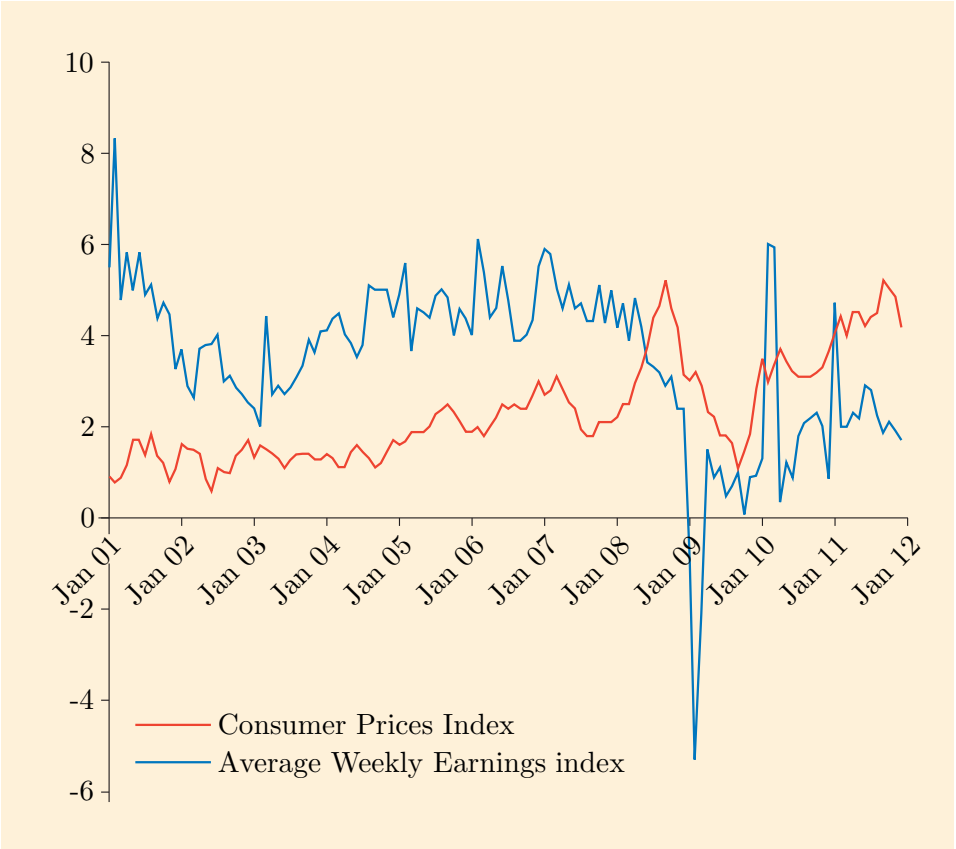


Figure 25 Changes in the CPI and the AWE index, 2001–2011 (percentage increases over previous year)

On the basis of these data, for people in the UK from 2009 to 2011 the 'average' answer to the question *Are people getting better or worse off?* is pretty clearly *Worse off!*.

The overall point here is essentially Mr Micawber's point. If, over the last year, your pay has gone up by more than prices went up, you can now buy everything you could buy a year ago, and more, and you are better off. If instead your pay has gone up by less than prices went up, you can no longer buy everything you could buy a year ago and you are worse off.

Example 8 Comparing the changes in two indices

The values of the CPI in January 2011 and January 2012 were 116.9 and 121.1 respectively. Thus its value in January 2012 as a percentage of its value one year earlier was

$$\frac{121.1}{116.9} \times 100\% = 103.592\,81\% \simeq 103.6\%.$$

We can now compare the changes in the two indices, calculated here and in Example 7. The AWE index in January 2012 was 100.5% of its value one year earlier. Thus earnings had increased (slightly), but over the same year, prices (as measured by the CPI) had increased more. So, in terms of these index numbers, on average people are worse off.

But *how much* worse off? To get a single numerical measure, rather than looking at the AWE index and CPI changes separately, it would be helpful to combine these two ratios somehow. As both these indices are ratios, the appropriate way to combine them is to calculate the ratio of the ratios. So we calculate the

following ratio (using the unrounded values):

$$\frac{100.483\,76}{103.592\,81}.$$

Expressing the answer as a percentage:

$$\frac{100.483\,76}{103.592\,81} \times 100\% = 96.998\,78\% \simeq 97.0\%.$$

Comparisons like the one in Example 8 give us a measure which is called the **real earnings for that month compared with one year earlier**. We would say that real earnings in January 2012 were 97.0% of their value a year earlier. That is, in terms of what they can buy, earnings in January 2012 were on average lower (indeed 3% lower) than they were one year earlier. We use the term 'real earnings' here because we take the ratio of actual earnings, given by the AWE index for that month divided by the AWE index one year earlier, and divide it by the corresponding ratio of price changes, obtained from the CPI values.

To calculate this measure directly, it can help to do a bit of rearranging. In Example 8, the ratio for comparing the CPI values over the year was

$$\frac{\text{CPI for January 2012}}{\text{CPI for one year earlier}}.$$

The corresponding ratio for comparing the AWE index values, from Example 7, was

$$\frac{\text{AWE index for January 2012}}{\text{AWE index for one year earlier}}.$$

Then, to get the real earnings measure, we divided the ratio calculated from the AWE index values by the corresponding ratio for the CPI, to give

$$\frac{\text{AWE index for January 2012}}{\text{AWE index for one year earlier}} \Big/ \frac{\text{CPI for January 2012}}{\text{CPI for one year earlier}},$$

which can be rearranged (using the usual rules for arithmetic with fractions) to

$$\frac{\text{AWE index for January 2012}}{\text{AWE index for one year earlier}} \times \frac{\text{CPI for one year earlier}}{\text{CPI for January 2012}}.$$

So, in general, we use the following formula.

Real earnings for month A compared with one year earlier

The real earnings for month A compared with one year earlier is defined as:

$$\frac{\text{AWE index for month A}}{\text{AWE index for one year earlier}} \times \frac{\text{CPI for one year earlier}}{\text{CPI for month A}}.$$

To illustrate this calculation we shall use the list of the values of both the CPI and the AWE index in Table 24.

Table 24 Values of the CPI and AWE index in 2010 and 2011

	CPI		AWE index	
	2010	2011	2010	2011
January	112.4	116.9	138.1	144.7
February	112.9	117.8	141.2	144.0
March	113.5	118.1	142.8	145.7
April	114.2	119.3	141.2	144.4
May	114.4	119.5	141.7	144.8
June	114.6	119.4	141.3	145.4
July	114.3	119.4	141.8	145.9
August	114.9	120.1	142.5	145.7
September	114.9	120.9	143.1	145.8
October	115.2	121.0	143.3	146.4
November	115.6	121.2	143.3	146.0
December	116.8	121.7	143.4	145.8

Example 9 Annual change in real earnings

The real earnings for December 2011 compared with one year earlier is:

$$\frac{\text{AWE index for December 2011}}{\text{AWE index for December 2010}} \times \frac{\text{CPI for December 2010}}{\text{CPI for December 2011}}$$

$$= \frac{145.8}{143.4} \times \frac{116.8}{121.7}$$

$$= 0.9757996 \simeq 0.976 = 97.6\%.$$

So the real earnings for December 2011 were 97.6% of their value a year earlier.

Activity 23 Annual changes at three time points

For each of the following months calculate, as a percentage, the real earnings for that month compared with one year earlier. (Round your answers to one decimal place.)

(a) March 2011 (b) June 2011 (c) September 2011

The calculation from Activity 23 has been done for every month in 2011, with Table 25 (below) showing the results (rounded to one decimal place).

Table 25 Real earnings in 2011 compared with one year earlier

	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec
%	100.7	97.7	98.1	97.9	97.8	98.8	98.5	97.8	96.8	97.3	97.2	97.6

There appears to be no obvious trend in these figures. The January figure shows a small increase in real earnings compared to the previous year (which was due largely to an increase in annual bonuses compared to the year before). Apart from that, every month shows a decrease in real earnings compared to the previous year.

5.3 Points to consider when using the AWE

The AWE index, like all statistics, can be misused. It cannot provide information for which it was not designed and so it must be used with care and common sense.

To conclude this section and the unit, you are asked to consider four points and to think how they could influence the use of the AWE index.

Averages are not individuals

The AWE index, like the CPI and RPI, is based on *averages* and should *not* be taken as representing the circumstances of any *particular individual* or individuals. So the AWE index and the CPI, even when used together, provide only a very poor assessment of whether a *particular group* of people is getting better or worse off.

Average means mean

In Subsection 5.1, the word *average* in the AWE index is always interpreted as *mean* or *weighted mean*. These both depend on the sum of the values in a particular batch. Remember that the formula is

$$\text{mean} = \frac{\text{sum}}{\text{size}}.$$

If the sum and the size remain the same, then the mean is unchanged, no matter how the individual values vary within the batch.

For example, suppose that Pat Gradgrind, the managing director of Gradgrind Ltd, receives a pay *increase* of *$*10 000 per year, and that each of ten cleaners in the firm gets a *decrease* of *$*1000 per year. Think for a few moments what effect this might have on the AWE index.

Changes such as that above will not affect the value of the AWE index even if Gradgrind is included in the Monthly Wages and Salaries Survey on which the AWE index is based. This is because the *total* amount paid (the sum) has not been changed, nor has the number of employees (the size). In short, the AWE index is not sensitive to such changes in the overall distribution of earnings – it is concerned only with *averages* (which, here, means *means*).

The AWE index and unemployment

A further problem connected with the AWE index can also be illustrated by events at Gradgrind Ltd. Suppose now that, instead of losing some earnings, the cleaners are all made redundant. What do you think the effect of this on the AWE index would be? (Not to mention the effect on the cleanliness of Gradgrind's premises!)

Cleaners are likely to be among the lowest-paid workers at Gradgrind, so if the cleaners are laid off, then the average pay of those remaining will go up. This means that the AWE index will increase! This may seem paradoxical, and is indeed difficult to understand given the feeling that if the AWE index is *increasing* then people are *getting better off*.

However, the paradox is resolved when we realise that the AWE index is based *only on those in employment* so it can increase purely as a result of lower-paid workers being laid off.

The impact of income tax

In the UK, income tax is charged on all income over a certain amount each year. For example, suppose that a typical UK taxpayer paid no income tax on the first *$*8000 of earnings but paid tax at the rate of 20% on all income above this. (We shall use these numbers here just to illustrate the points we are making. In practice, in the UK the *actual* numbers change every year, or thereabouts, with the changes generally being announced in the Budget statement.)

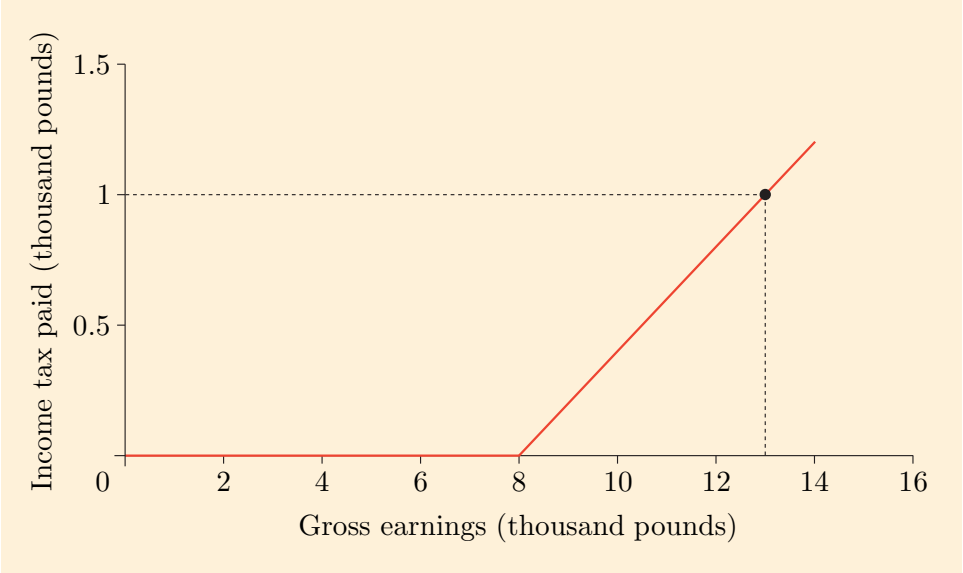


Figure 26 Relationship between income tax and gross earnings

This link between earnings and tax payable is shown in Figure 26. For example, an individual earning *$*13 000 each year pays tax on the earnings above *$*8000. This means a 20% charge on *$*5000 – that is, a tax bill of *$*1000 for that year.

Has this anything to do with the AWE index? Yes, it has, because it affects how much *earnings* must increase to compensate for a given increase in *prices*.

Suppose further that, as measured by the CPI, *prices* increase by 10%, and that the AWE index also increases by 10%. On the surface, it would appear that, on average, people would be just as well off as before – however, that is not the case.

To illustrate the problem, consider Martha Greenslade, one of the cleaners at Gradgrind Ltd, who earns exactly *$*8000 per year. Martha therefore pays no tax. Now suppose that Martha is an 'average' person in the sense that she gets a pay increase equal to the increase in the AWE index, namely 10%. In other words, Martha's earnings increase from *$*8000 to *$*8800 per year.

However, these are *gross* earnings. Martha has now gone into the 20% tax bracket, so she must pay this rate of tax on the whole of her increase. Therefore the increase in her *net* earnings is not *$*800, but only *$*640. (*Net income* is gross income minus the amount of tax paid.) In other words, an increase of 10% in her *gross* earnings has been transformed by taxation into an increase of only 8% in her *net* earnings. If the CPI went up by 10% and if Martha wanted to match this by a 10% increase in her *net* earnings (after all, it is her *net* earnings out of which she must pay the higher prices) then the increase in her *gross* earnings would have to be 12.5%, since this is the increase which reduces to 10% after allowing for taxation.

Of course, different people are affected in different ways by taxation, and Martha is a rather artificial example, being just on the edge of the 20% tax bracket. It is

true that for earnings after taxation to match inflation (the rise in prices) it is necessary for the AWE index to increase by rather more than the CPI. In practice, however, UK governments normally increase tax-free allowances annually in line with inflation so that the proportion of taxable earnings remains roughly the same.

Activity 24 Changes in Martha's net pay

For each of the following two amounts, suppose this was Martha's earnings before receiving the 10% increase in her gross pay. What would be the percentage increase in her net pay?

(a) *$*7500

(b) *$*8500

A clear message of Subsection 5.3 has been that making judgements about people's earnings based solely on averages like the AWE index can be very misleading. Indeed, the same warnings need to be applied to all available average measures of earnings and price changes – they draw only on data taken from an incomplete population and they say nothing about the inequalities experienced by individuals. Taken overall, the increases of average earnings did seem to have exceeded price rises from 2001 to about the end of 2007, although from then until the time of writing (2012) at least, the position is reversed and price rises have generally been exceeding rises in average earnings. So, are people getting better or worse off? In an important sense it depends on the timescale. Since 2008 up to the time of writing, on average people are *not* getting better off. But, for instance, AWE index and CPI figures can be used to calculate the real earnings in January 2012 as a percentage of the real earnings in January 2000. In terms of that measure, real earnings increased by 12.6% over that twelve-year period (even though it includes the period since 2008 during which real earnings were decreasing). However, earnings inequalities have widened since 2000 (and indeed for some time before that). So, compared to the position of the highest earners, the position of those on low pay has got worse. We have some partial answers to the question *Are people getting better or worse off?*, but in some ways the main thing that we have learned is that this question has no straightforward answer.

Summary

In this unit you have learned how statistics can answer questions connected with people's earnings. You have learned how to calculate earnings ratios at different points across distributions. These ratios summarise the pay differential between men's and women's earnings. They can depend on factors such as the inclusion or exclusion of overtime, the numbers of hours worked, the sector (public or private) and the occupation.

You have learned more about boxplots, including how to recognise skewness in a data from a boxplot and how to construct boxplots, both by hand and by using Minitab. In ordinary boxplots, the ends of the whiskers represent the upper and lower adjacent values, whereas, in decile boxplots, the ends of the whiskers represent the highest and lowest deciles.

A new measure of spread – the standard deviation – has been described. The standard deviation is paired with the mean, whereas the interquartile range is

paired with the median. You learned two methods to calculate the standard deviation by hand, along with how to obtain the standard deviation and other summary measures, using Minitab. You also learned how to calculate the mean and standard deviation by hand for grouped data.

The Average Weekly Earnings index is published by the Office for National Statistics and measures the changes in earnings. By comparing this index with the CPI, you have been able to calculate real earnings compared with a year before and hence go some way to answering the question: *Are people getting better or worse off?*

Finally, you have been introduced to some surveys which collect data on earnings in the UK. You have seen how the way the data are collected can limit the conclusions which can be drawn from an analysis.

Learning outcomes

After working through this unit, you should be able to:

- calculate earnings ratios and understand how they measure the discrepancy between men's and women's earnings
- describe briefly the Annual Survey of Hours and Earnings (ASHE) and some of the data which it collects
- use the median, quartiles, and highest and lowest deciles to describe large batches of data
- interpret decile boxplots
- understand that, in order to compare like with like, it is essential to select from published sources of data with great care
- interpret and compare boxplots in terms of skewness
- draw a boxplot, dealing appropriately with adjacent values and potential outliers
- calculate the standard deviation of a batch of data
- calculate the mean and standard deviation for a batch of grouped data
- understand some of the factors affecting the choice of summary measures for a batch of data
- use Minitab to do further numerical calculations on data
- use Minitab to obtain and customise boxplots
- describe what is meant in the Average Weekly Earnings (AWE) by earnings
- describe the main stages involved in the calculation of the AWE index
- compare changes in the AWE index with changes in the CPI
- calculate real earnings compared with a year earlier
- describe how the interpretation of the AWE index is affected by:
  - the distribution of earnings
  - unemployment
  - taxation.

Solutions to activities

Solution to Activity 1

(a) Women worked 37.4 hours per week on average. This is fewer hours than the average worked by men, which was 40.2 hours.

(b) On average, men did 1.0 hours more overtime per week than women $(1.5 - 0.5 = 1.0)$. Alternatively, men did three times as much overtime as women $(1.5/0.5 = 3)$.

(c) Removing overtime pay from the gross weekly earnings figures would reduce the men's figure more than the women's figure, since men did more overtime. You would expect this to narrow the 'gap' between men's earnings and women's earnings and therefore *increase* the earnings ratio.

(d) Since men worked more hours per week on average than women, you would expect men's gross weekly earnings to be more than women's, even if they were paid the same for similar amounts of work. A fairer comparison might be to look at the gross *hourly* earnings of men and women. This would eliminate the effect on earnings of men working more hours per week than women.

Solution to Activity 2

(a) The earnings ratio at the mean based on weekly earnings excluding overtime is

$$\frac{509}{635} \simeq 0.801\,57 \text{ or } 80\%.$$

(b) The earnings ratio at the mean based on hourly earnings is

$$\frac{1382}{1643} \simeq 0.841\,14 \text{ or } 84\%.$$

(c) Removing overtime pay from gross weekly earnings increases the earnings ratio at the mean from 78% to 80%. Comparing hourly earnings instead of weekly earnings increases the earnings ratio at the mean further to 84%.

Solution to Activity 3

(a) The earnings ratio at the median for gross weekly earnings including overtime is calculated as

$$\frac{440}{538} \simeq 0.817\,84,$$

or 82% after rounding to the nearest one per cent.

The other earnings ratios at the median are calculated similarly, and all of them are given in the table below.

Earnings ratios at the median	%
Gross weekly earnings including overtime	82
Gross weekly earnings excluding overtime	85
Gross hourly earnings excluding overtime	90

(b) As was the case when using the mean, the earnings ratio at the median increases when overtime is excluded and again when hourly earnings are considered instead of weekly earnings. In each case, the earnings ratio at the median is higher than the corresponding earnings ratio at the mean.

Solution to Activity 4

(a) As, by definition, 50% of people earn less than the median wage, a person of median earnings will pass by halfway through the parade at 10:30 am.

(b) You were told that a person of mean height passes by 25 minutes before the end of the parade – that is, at 10:35 am. This is later than the time a person of median earning passes, and reflects the fact that the mean earnings is greater than the median earnings – as is generally the case for right-skew data like these.

Because 10:35 am is 35/60 of the way through the hour, 35/60 or 58% of people earn less than the mean wage. (Actually, the proportion is over 60% as the precise time that a person of average height would pass is actually about 23 minutes before the end, not 25.)

Solution to Activity 5

See text below the activity for discussion of this.

Solution to Activity 6

(a) The 25th percentile has 25%, that is, one quarter, of the batch below it. The 75th percentile has three-quarters of the batch below it, so one quarter of the batch above it. Therefore these two percentiles are actually the quartiles. The 25th percentile is the lower quartile, and the 75th percentile is the upper quartile.

(b) The 50th percentile has 50%, that is, half, the batch below it, and therefore half the batch above it. Thus it is the median, so from the table the 50th percentile is *$*432 for women and *$*509 for men.

Solution to Activity 7

(a) 1083 is the highest decile for the group, so 10% of the men earned more than *$*1083.

(b) 284 and 509 are, respectively, the lowest decile and the median, so 40% of these men earned between *$*284 and *$*509.

(c) 364 is the lower quartile, so 25% of the 10 652 000 men, or about 2 663 000 men, earned less than *$*364.

Solution to Activity 8

The earnings ratios for the upper quartile is

$$\frac{619}{738} = 0.838\,753,$$

which is 84% to the nearest one per cent. The other earnings ratios, calculated similarly, are 87% at the lower quartile, 76% at the highest decile and 89% at the lowest decile. (Rounded from 86.8132%, 75.7156% and 89.0845%.)

Solution to Activity 9

(a) The following table shows the earnings ratio at the median for each year.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Earnings ratio (%)	83	83	84	84	84	85	85	85

Year	2005	2006	2007	2008	2009	2010	2011
Earnings ratio (%)	87	87	88	87	88	90	90

(b) There was a slow, steady increase in the earnings ratio over this period.

(c) Since the earnings ratio has increased, this measure suggests gender inequalities in earnings have narrowed.

Solution to Activity 10

Here are the earnings ratios (in percentages) for each group. The all-workers earnings ratios are also given again, so as to make comparisons easier.

	Public	Private	All
Highest decile	79	72	76
Upper quartile	86	77	84
Median	87	78	85
Lower quartile	88	81	87
Lowest decile	90	86	89

The earnings ratios for public sector workers are higher than those for all workers, and the ratios for private sector workers are noticeably lower than those for all workers. Therefore, women's earnings seem to be nearer to men's earnings in the public sector than in the private sector, with the earnings ratios calculated across all workers being somewhere in between.

However, the earnings ratios for all workers together are closer to those for the public sector than those for the private sector, throughout the range, but particularly towards the lower end.

Solution to Activity 11

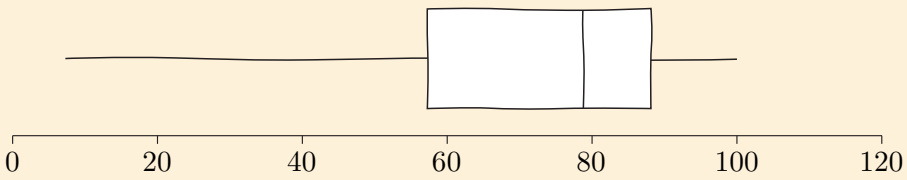
The earnings ratios shown below suggest that there are clear differences between occupations. For example, women who are managers and directors in retail and wholesale seem particularly underpaid, especially at the middle and top of the earnings scale, whereas in secondary education, the gender gap is relatively narrow, and it is almost (but not quite!) non-existent for kitchen and catering assistants.

Occupation	Sales and retail assistants			Secondary education professionals			Kitchen and catering assistants			Managers and directors in retail/wholesale		
	M	W	R	M	W	R	M	W	R	M	W	R
Upper quartile	352	312	89	852	801	94	297	283	95	698	497	71
Median	289	260	90	737	699	95	247	245	99	502	374	75
Lower quartile	248	227	92	624	560	90	216	213	99	383	302	79

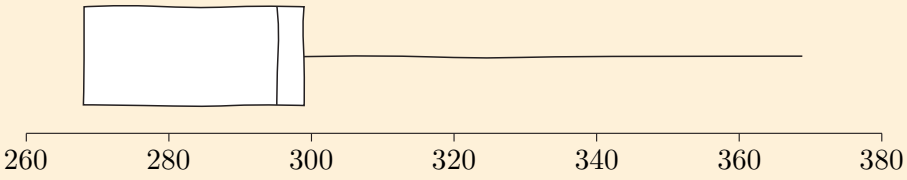
M: Men W: Women R: Earnings ratio

Solution to Activity 12

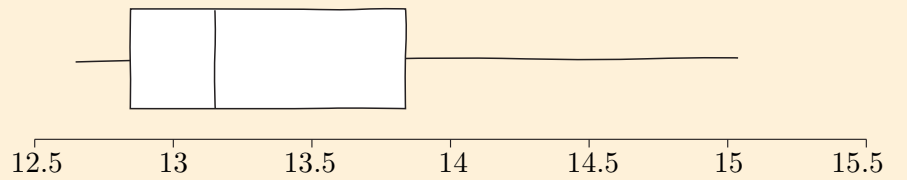
(a) The boxplot for the arithmetic scores is shown below. Here the left whisker of the boxplot is much longer than the right. Also, the left part of the box is much longer than the right part. Both of these characteristics indicate that this batch is left-skew.



(b) The boxplot for the coffee prices is shown below. In this case, the right whisker of the boxplot is much longer than the left (which has zero length because the lower quartile is equal to the lower extreme). This indicates right skewness. However, the left part of the box is much longer than the right part, indicating left skewness. It would therefore be misleading to describe this batch as either left-skew or right-skew. All you can really say is that it is certainly not symmetric.



(c) The boxplot for the electricity prices is shown below. In this case, the right whisker of the boxplot is much longer than the left and the right part of the box is longer than the left part. Both these characteristics indicate that the batch is right-skew.



Solution to Activity 13

(a) For the five-figure summary, you need the extremes, the quartiles, and the median.

The minimum is 6, and the maximum is 43.

For a batch size of 40, the median position is $\frac{1}{2}(40+1) = 20\frac{1}{2}$. The median is therefore halfway between the 20th and the 21st values in order. These are both 10, so the median is 10.

The quartile positions are $\frac{1}{4}(40+1) = 10\frac{1}{4}$ and $\frac{3}{4}(40+1) = 30\frac{3}{4}$. So Q_1 is one quarter of the way from $x_{(10)}$ to $x_{(11)}$ (where, as in Unit 2, $x_{(10)}$ means the 10th value in order, and so on). Here $x_{(10)} = x_{(11)} = 8$, so $Q_1 = 8$. Then Q_3 is three quarters of the way from $x_{(30)}$ to $x_{(31)}$, and here $x_{(30)} = 15$ and $x_{(31)} = 17$. The difference between these two values is 2, and three quarters of 2 is 1.5, so the upper quartile is $Q_3 = 15 + 1.5 = 16.5$, but (again, as usual in Unit 2), for presenting in a five-figure summary this should be rounded to the accuracy of the original data: $Q_3 \simeq 17$.

The five-figure summary is therefore as follows.

		10	
$n = 40$	8		17
	6		43

(b) The only extra calculations for drawing a boxplot are to find the interquartile range and hence the adjacent values. Remember to use unrounded numbers in intermediate calculations, and only round at the end.

$$\text{IQR} = Q_3 - Q_1 = 16.5 - 8 = 8.5$$

and so

$$Q_1 - 1.5 \times \text{IQR} = 8 - 1.5 \times 8.5 = 8 - 12.75 = -4.75.$$

The smallest data value that is not less than this is the minimum, 6, so the lower adjacent value is 6, and hence the lower whisker will go down from the lower quartile, 8, to 6.

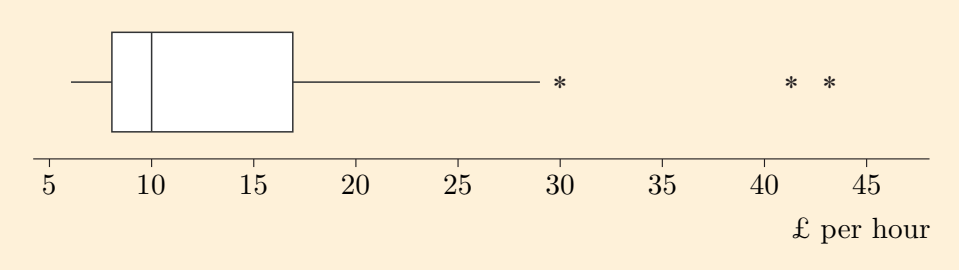
$$Q_3 + 1.5 \times \text{IQR} = 16.5 + 1.5 \times 8.5 = 16.5 + 12.75 = 29.25.$$

The highest data value that does not exceed 29.25 is 29, so the upper adjacent value is 29. The upper whisker will go up from the upper quartile, 16.5, to 29.

There are three data values, 30, 41 and 43, that are not covered by the whiskers – these are potential (high) outliers and should be plotted separately.

In drawing the boxplot, the scale has to go at least as far as from one extreme to the other, that is, from 6 to 43. One possibility would be to have a scale running from 0 to 45, with the axis marked every $5, but there are other reasonable possibilities.

The resulting plot should look like the one below. (Yours might have a different scale marked on the axis.)



Boxplot of earnings for 40 female employees

Solution to Activity 14

(a) For the programmers:

$$\begin{aligned}\sum x &= 465 + 484 + 620 + 654 + 855 + 858 \\ &= 3936\end{aligned}$$

and the mean is

$$\overline{x} = \frac{\sum x}{n} = \frac{3936}{6} = 656.$$

For the others:

$$\begin{aligned}\sum x =& 346 + 376 + 391 + 391 + 415 + 465 + 830 \\ &+ 843 + 876 + 1627 \\ =& 6560\end{aligned}$$

and the mean is

$$\overline{x} = \frac{\sum x}{n} = \frac{6560}{10} = 656.$$

(b) The first batch size is 6, so the median lies halfway between the third and fourth value. These are 620 and 654, so the median is 637.

The second batch size is 10, so the median lies halfway between the fifth and sixth values. These are 415 and 465, so the median is 440.

(c) For the first batch, the positions of the quartiles are $\frac{1}{4}(6+1) = 1.75$ and $\frac{3}{4}(6+1) = 5.25$. Thus the lower quartile is three quarters of the way from the first value to the second value, so it is

$$Q_1 = 465 + \tfrac{3}{4}(484 - 465) = 479.25.$$

The upper quartile is one quarter of the way from the fifth value to the sixth value, so it is

$$Q_3 = 855 + \tfrac{1}{4}(858 - 855) = 855.75.$$

Thus

$$\text{IQR} = Q_3 - Q_1 = 855.75 - 479.25 = 376.5 \simeq 377.$$

For the 'others' batch, the positions of the quartiles are $\frac{1}{4}(10+1) = 2.75$ and $\frac{3}{4}(10+1) = 8.25$. Thus the lower quartile is three quarters of the way from the second value to the third value, so it is

$$Q_1 = 376 + \tfrac{3}{4}(391 - 376) = 387.25.$$

The upper quartile is one quarter of the way from the eighth value to the ninth value, so it is

$$Q_3 = 843 + \tfrac{1}{4}(876 - 843) = 851.25.$$

Thus

$$\text{IQR} = Q_3 - Q_1 = 851.25 - 387.25 = 464.$$

Solution to Activity 15

The completed table is as follows.

Data x	Mean $\overline{x}$	Deviation $(x-\overline{x})$	Squared deviation $(x-\overline{x})^2$
465	656	-191	36 481
484	656	-172	29 584
620	656	-36	1 296
654	656	-2	4
855	656	199	39 601
858	656	202	40 804
	$\sum$	0	147 770

Solution to Activity 16

The completed table is as follows.

Data x	Mean $\overline{x}$	Deviation $(x-\overline{x})$	Squared deviation $(x-\overline{x})^2$
346	656	-310	96 100
376	656	-280	78 400
391	656	-265	70 225
391	656	-265	70 225
415	656	-241	58 081
465	656	-191	36 481
830	656	174	30 276
843	656	187	34 969
876	656	220	48 400
1627	656	971	942 841
	$\sum$	0	1 465 998

The variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{1\,465\,998}{9} = 162\,888.667.$$

The standard deviation is

$$\sqrt{162\,888.667} \simeq 403.6.$$

Solution to Activity 17

The completed table is as follows.

	Data x	Squared data x^2
	346	119 716
	376	141 376
	391	152 881
	391	152 881
	415	172 225
	465	216 225
	830	688 900
	843	710 649
	876	767 376
	1627	2 647 129
$\sum$	6560	5 769 358

Then the sum of the squared deviations is

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &= 5\,769\,358 - \frac{6560^2}{10}\\ &= 5\,769\,358 - \frac{43\,033\,600}{10}\\ &= 5\,769\,358 - 4\,303\,360 = 1\,465\,998.\end{aligned}$$

This is exactly the result you found for the sum of the squared deviations, $\sum(x-\overline{x})^2$, in Activity 16. As found in that activity, the variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{1\,465\,998}{9} = 162\,888.667,$$

and the standard deviation is $\sqrt{162\,888.667} \simeq 403.6$.

Solution to Activity 18

To use Method 2 we need to know, for the combined batch, both of the following sums: $\sum x$ and $\sum x^2$.

These can both be obtained easily by adding the corresponding sums for the two separate batches in Table 16 and in the solution to Activity 17. Thus, for the combined batch, we have

$$\sum x = 3936 + 6560 = 10\,496,$$

$$\sum x^2 = 2\,729\,786 + 5\,769\,358 = 8\,499\,144,$$

and the size of the combined batch is $n = 6 + 10 = 16$.

Then, the mean is

$$\frac{\sum x}{n} = \frac{10\,496}{16} = 656.$$

Next, the sum of the squared deviations is

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &= 8\,499\,144 - \frac{(10\,496)^2}{16}\\ &= 8\,499\,144 - \frac{110\,166\,016}{16}\\ &= 8\,499\,144 - 6\,885\,376 = 1\,613\,768.\end{aligned}$$

So the variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{1\,613\,768}{15} = 107\,584.533,$$

and the standard deviation is $\sqrt{107\,584.533} = 328.0$ (rounded to one decimal place).

Solution to Activity 19

The total number of red blood cells is $n = \sum f = 50\,000$.

The sum of the squared deviations is

$$\begin{aligned}\sum (x - \overline{x})^2 f &= \sum x^2 f - \frac{(\sum xf)^2}{n} \\ &= 14\,884 - \frac{11\,520^2}{50\,000} \\ &= 14\,884 - 2654.208 = 12\,229.792.\end{aligned}$$

Then,

$$\text{variance} = \frac{12\,229.792}{n-1} = \frac{12\,229.792}{49\,999} = 0.244\,6007.$$

The standard deviation is the square root of the variance, so it is

$$s = \sqrt{0.244\,6007} = 0.494\,5713 \simeq 0.49.$$

Solution to Activity 20

The required sums can be calculated as in the following table.

x	x^2	f	xf	$x^2 f$
0	0	229	0	0
1	1	211	211	211
2	4	93	186	372
3	9	35	105	315
4	16	7	28	112
5	25	1	5	25
	$\sum$	576	535	1035

The mean is

$$\frac{\sum xf}{\sum f} = \frac{535}{576} = 0.928\,819 \simeq 0.9.$$

The sum of the squared deviations is

$$\begin{aligned}\sum (x - \overline{x})^2 &= \sum x^2 f - \frac{(\sum xf)^2}{\sum f} \\ &= 1035 - \frac{535^2}{576} \\ &= 1035 - 496.918\,403 = 538.081\,597.\end{aligned}$$

To find the variance, divide by one less than the batch size:

$$\text{variance} = \frac{538.081\,597}{575} = 0.935\,794.$$

The standard deviation is the square root of the variance, so it is

$$s = \sqrt{0.935\,794} = 0.967\,365 \simeq 1.0.$$

Solution to Activity 21

The quartiles were between the second and third values, and the eighth and ninth values, respectively, for this batch. None of these values have changed, so the quartiles have not changed either. Also, as the quartiles have not changed, nor has the interquartile range: it is still 464.

However, we must recalculate the standard deviation.

	Data x	Squared data x^2
	246	60 516
	376	141 376
	391	152 881
	391	152 881
	415	172 225
	465	216 225
	830	688 900
	843	710 649
	876	767 376
	1727	2 982 529
$\sum$	6560	6 045 558

The sum of the squared deviations is now

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum(x^2) - \frac{(\sum x)^2}{n}\\ &= 6\,045\,558 - \frac{6560^2}{10}\\ &= 6\,045\,558 - \frac{43\,033\,600}{10}\\ &= 6\,045\,558 - 4\,303\,360 = 1\,742\,198.\end{aligned}$$

The variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{1\,742\,198}{9} = 193\,577.556,$$

and the standard deviation is

$$\sqrt{193\,577.556} \simeq 440.0.$$

Solution to Activity 22

The value of the index in January 2001 as a percentage of its January 2000 value was

$$\frac{103.6}{98.2} \times 100\% = 105.498\,98\% \simeq 105.5\%.$$

Thus its increase over the year January 2000 to January 2001 was 5.5% of its January 2000 value.

Solution to Activity 23

(a) Real earnings for March 2011 compared with one year earlier is:

$$\frac{\text{AWE index for March 2011}}{\text{AWE index for March 2010}} \times \frac{\text{CPI for March 2010}}{\text{CPI for March 2011}}$$
$$= \frac{145.7}{142.8} \times \frac{113.5}{118.1}$$
$$= 0.980\,5671 \simeq 0.981 = 98.1\%.$$

(b) Real earnings for June 2011 compared with one year earlier is:

$$\frac{\text{AWE index for June 2011}}{\text{AWE index for June 2010}} \times \frac{\text{CPI for June 2010}}{\text{CPI for June 2011}}$$
$$= \frac{145.4}{141.3} \times \frac{114.6}{119.4}$$
$$= 0.987\,6488 \simeq 0.988 = 98.8\%.$$

(c) Real earnings for September 2011 compared with one year earlier is:

$$\frac{\text{AWE index for September 2011}}{\text{AWE index for September 2010}} \times \frac{\text{CPI for September 2010}}{\text{CPI for September 2011}}$$
$$= \frac{145.8}{143.1} \times \frac{114.9}{120.9}$$
$$= 0.968\,3038 \simeq 0.968 = 96.8\%.$$

Solution to Activity 24

(a) With earnings of *$*7500, Martha would not pay any tax and so her net pay would be *$*7500 before the increase. After receiving a 10% increase on *$*7500, Martha would receive

$$\$7500 + \$750 = \$8250.$$

Tax on this would be 20% of *$*250, which equals *$*50. Therefore her net income would be *$*8200, which is a 9.3% increase on her previous net pay (rounded to one decimal place).

(b) With earnings of *$*8500, Martha would pay *$*100 tax and receive *$*8400 as net pay before the increase. After the increase her gross pay would be

$$\$8500 + \$850 = \$9350.$$

Tax on this would be 20% of *$*1350, which equals *$*270. Therefore her net pay after the increase would be

$$\$9350 - \$270 = \$9080.$$

This is an 8.1% increase on her previous net pay (rounded to one decimal place).

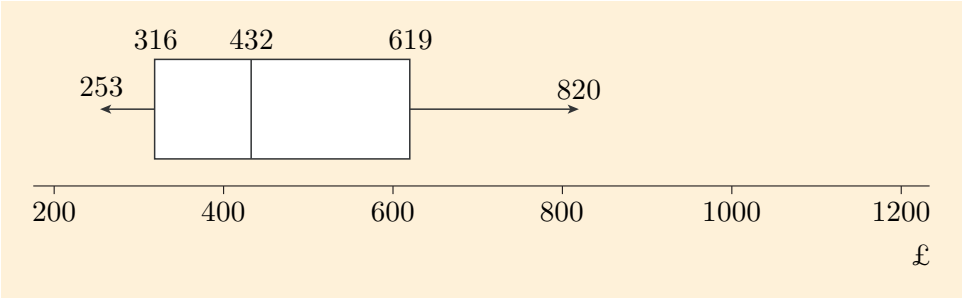
Solutions to exercises

Solution to Exercise 1

813 is the 80th percentile, so 80% of the men earned less than $813, and so 20% of the 10 652 000 men, that is, about 2 130 400 men, earned $813 or more.

Solution to Exercise 2

Here is the decile boxplot from Figure 10 with the key numerical values positioned.



Solution to Exercise 3

As a percentage, the earnings ratio at the highest decile, for example, is

$$\frac{767}{966} \times 100\% = 79.399\,586\% = 79\% \text{ (to the nearest one per cent).}$$

Calculating the other ratios in a similar way gives the following list.

Percentile	Earnings ratio
Highest decile	79%
Upper quartile	83%
Median	83%
Lower quartile	81%
Lowest decile	87%

The earnings ratio is noticeably higher at the bottom end of the distribution than at the top end. The ratios are fairly similar to those for jobs in all industries, as given in Table 8 (in Subsection 1.6).

Solution to Exercise 4

(a) For a batch size of 35, the median position is $\frac{1}{2}(35+1) = 18$. So the median is the 18th data value in order, which is 12. The quartiles are at positions $\frac{1}{4}(35+1) = 9$ and $\frac{3}{4}(35+1) = 27$. Therefore $Q_1 = 8$ and $Q_3 = 17$.

(b) Here IQR $= 17 - 8 = 9$. So

$$Q_1 - 1.5 \times \text{IQR} = 8 - 1.5 \times 9 = 8 - 13.5 = -5.5.$$

There are no data values in Figure 19 less than -5.5 so the lower adjacent value is actually the minimum of the batch, 6.

Also,

$$Q_3 + 1.5 \times \text{IQR} = 17 + 1.5 \times 9 = 17 + 13.5 = 30.5.$$

The highest data value that does not exceed 30.5 is 29, so the upper adjacent value is 29.

There is only one data value, 38, that is not covered by the whiskers (that is, not between the adjacent values), so only 38 should be plotted separately.

Solution to Exercise 5

(a) For a batch size of 34, the median position is $\frac{1}{2}(34+1) = 17.5$. So the median is halfway between the 17th and 18th data values in order. Since these values are both 35.3, the median is 35.3.

The quartiles are at positions $\frac{1}{4}(34+1) = 8\frac{3}{4}$ and $\frac{3}{4}(34+1) = 26\frac{1}{4}$.

Therefore, the lower quartile is three quarters of the way between the eighth and ninth values in order, which are 30.1 and 31.7. The difference between these values is $31.7 - 30.1 = 1.6$, so

$$Q_1 = 30.1 + \tfrac{3}{4} \times 1.6 = 31.3.$$

The upper quartile is a quarter of the way between the 26th and 27th values in order, which are 36.7 and 37.5. The difference between these values is $37.5 - 36.7 = 0.8$, so

$$Q_3 = 36.7 + \tfrac{1}{4} \times 0.8 = 36.9.$$

(b) Here IQR $= 36.9 - 31.3 = 5.6$. So

$$Q_1 - 1.5 \times \text{IQR} = 31.3 - 1.5 \times 5.6 = 31.3 - 8.4 = 22.9.$$

There are no data values in Figure 20 less than 22.9 so the lower adjacent value is the minimum of the batch, 26.1.

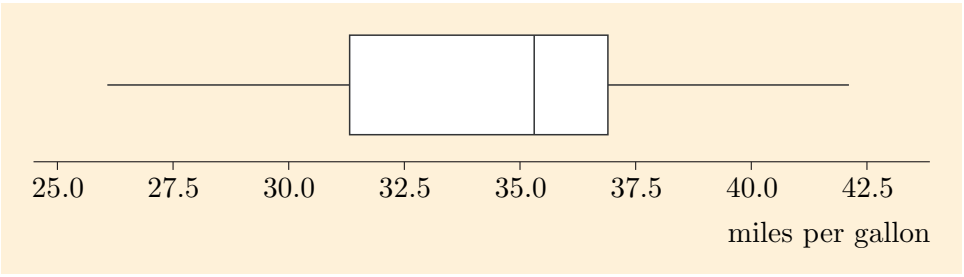
Also,

$$Q_3 + 1.5 \times \text{IQR} = 36.9 + 1.5 \times 5.6 = 36.9 + 8.4 = 45.3.$$

There are no data values in Figure 20 greater than 45.3 so the upper adjacent value is the maximum of the batch, 42.1.

In this case, no data values lie outside the adjacent values, so none are plotted separately.

(c) The boxplot is as follows. The two whiskers have the same length, suggesting that the data are symmetric. However, the left part of the box is longer which suggests that the data are left-skew. The general shape of the stemplot also indicates left skewness.



Solution to Exercise 6

For the 'Before insulation' batch:

x	x^2
12.1	146.41
11.0	121.00
14.1	198.81
13.8	190.44
15.5	240.25
12.2	148.84
12.8	163.84
9.9	98.01
10.8	116.64
12.7	161.29
$\sum$ 124.9	1585.53

$$\overline{x} = \frac{\sum x}{n} = \frac{124.9}{10} = 12.49.$$

$$\sum (x - \overline{x})^2 = 1585.53 - \frac{124.9^2}{10} = 25.529.$$

$$s = \sqrt{\frac{25.529}{9}} = \sqrt{2.836\,556} \simeq 1.68.$$

So the mean is 12.49 and the standard deviation is approximately 1.68.

For the 'After insulation' batch:

x	x^2
12.0	144.00
10.6	112.36
13.4	179.56
11.2	125.44
15.3	234.09
13.6	184.96
12.6	158.76
8.8	77.44
9.6	92.16
12.4	153.76
$\sum$ 119.5	1462.53

$$\overline{x} = \frac{\sum x}{n} = \frac{119.5}{10} = 11.95.$$

$$\sum (x - \overline{x})^2 = 1462.53 - \frac{119.5^2}{10} = 34.505.$$

$$s = \sqrt{\frac{34.505}{9}} = \sqrt{3.833\,889} \simeq 1.96.$$

So the mean is 11.95 and the standard deviation is approximately 1.96.

Solution to Exercise 7

The required sums can be calculated as in the following table.

x	x^2	f	xf	x^2f
0	0	27	0	0
1	1	44	44	44
2	4	26	52	104
3	9	3	9	27
	$\sum$	100	105	175

The mean is $\dfrac{105}{100} = 1.05$.

The sum of squared deviations is

$$\begin{aligned}\sum (x-\overline{x})^2 &= 175 - \frac{105^2}{100} \\ &= 175 - 110.25 = 64.75.\end{aligned}$$

To find the variance, divide by one less than the batch size:

$$\text{variance} = \frac{64.75}{99} = 0.654\,040.$$

The standard deviation is the square root of the variance, so it is

$$s = \sqrt{0.654\,040} = 0.808\,728 \simeq 0.81.$$

Acknowledgements

Grateful acknowledgement is made to the following sources:

Table 9 from the Office of National Statistics (2011) 'Patterns of Pay, 1997 to 2011 ASHE Results', reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Table 11 from the Office of National Statistics (2011) 'Patterns of Pay, Annual Survey of Hours and Earnings', reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Table in solution to Activity 11, from the Office of National Statistics (2011), 'Patterns of Pay, Annual Survey of Hours and Earnings', reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Table 18 Wang C.C. (1970) *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 64, Elsevier Limited

Table 21 Clarke R.D. (1946) 'An Application of the Poisson Distribution', *Journal of the Institute of Actuaries*, vol. 72

Table 23 Bailey B.J.R. (1990) *Applied Statistics*, vol. 39

Table 24 taken from the Office of National Statistics, 'CPI and AWE Index Values in 2010 and 2011', reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Figure 13 Rebecca Simpson / Dame Alice Owen's School, Potters Bar

Figure 17 taken from the Office of National Statistics, '2011 Annual Survey of Hours and Earnings'. Reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Figure 21 US Federal Government

Figure 22 Courtesy of Londonist. www.londonist.com/2009/01/london_v2_sitemapped.php.

Figure 24 CPI Moto Limited

Subsection 1.2 cartoon (gender scales), www.cartoonmovement.com

Subsection 1.4 cartoon (median height), taken from www.medicine.mcgill.ca/epidemiology/hanley/tmp/DescriptiveStatistics/median_or_mean_height.gif

Subsection 1.4 cartoon (pay parade), Jeremy Banx / Banxcartoons.co.uk

Subsection 3.1 cartoon, Randy Glasbergen, http://www.glasbergen.com

Subsection 5.1 quote from the Office of National Statistics, 'Average Weekly Earnings (AWE) Index'. Reproduced under the terms of the OGL, www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index